(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2003－122669
(P2003－122669A)
(43)公開日　平成15年４月25日(2003.4.25)

(51)Int.Cl.⁷		識別記号	FI		テーマコード(参考)
Ｇ０６Ｆ	13/00	５４０	Ｇ０６Ｆ	13/00	５４０Ｒ　５Ｂ０７５
	17/30	１１０		17/30	１１０Ｆ
		２２０			２２０Ｚ
		４１９			４１９Ｂ
	17/60	１４８		17/60	１４８

審査請求　未請求　請求項の数10　ＯＬ　（全 27 頁）

(21)出願番号　特願2001－314993(P2001－314993)
(22)出願日　平成13年10月12日(2001.10.12)

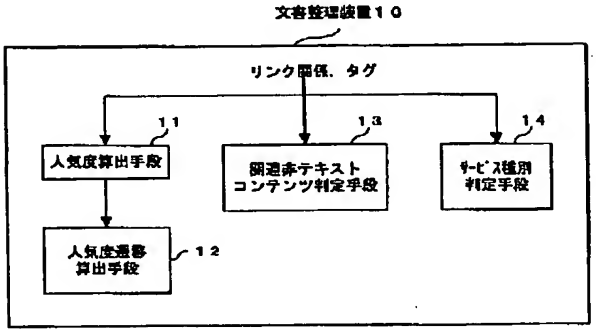
(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(72)発明者　津田　宏
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(74)代理人　100074099
弁理士　大菅　義之　（外１名）

Ｆターム(参考)　5B075 ND03 ND06 ND14 ND36 NR01 PQ40 PQ46

(54)【発明の名称】　リンク関係に基づく文書整理方法

(57)【要約】
【課題】　文書の人気度が増加する一方で減少することがないという問題を解決しつつ、文書が時系列的にどのような状況にあるのかを示す情報を得る事を可能とする。
【解決手段】　文書整理装置１０は、人気度算出手段１１及び人気度遷移算出sy段１２を備える。人気度算出手段１１は、第１の期間に収集されたネットワーク上の文書間のリンク関係に基づいて、各文書の人気の高さの度合いを示す人気度を算出する。人気度遷移算出手段１２は、第２の期間内に人気度算出手段１１によって算出された人気度に基づいて、人気度の変化の方向と度合いを示す人気変化度を算出する。

本　発　明　の　原　理　図



【特許請求の範囲】
【請求項１】　ネットワーク上の文書の人気の高さの度合いである人気度を算出する制御をコンピュータに実行させるプログラムであって、
文書からリンク関係を抽出し、
第１の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、
前記抽出された各文書の人気度を算出する、
ことを含む処理を前記コンピュータに実行させることを特徴とするプログラム。
【請求項２】　前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出する、
ことを更に含む処理を更にコンピュータに実行させることを特徴とする請求項１に記載のプログラム。
【請求項３】　第２の期間内に算出された前記人気度に基づいて、前記人気変化度を算出する、
ことを更に含む処理を前記コンピュータに実行させる事を特徴とする請求項２に記載のプログラム。
【請求項４】　前記第２の期間内に算出された前記人気度の時間に対する回帰式を算出し、
前記人気変化度を前記回帰式に基づいて算出する、
ことを更に含む処理を前記コンピュータに実行させる事を特徴とする請求項３に記載のプログラム。
【請求項５】　前記回帰式の回帰係数に基づいて前記人気変化度を決定する、
ことを更に含む処理を前記コンピュータに実行させる事を特徴とする請求項４に記載のプログラム。
【請求項６】　前記回帰式の切片に基づいて、前記人気度の時間に対する推移の傾向を決定する、
ことを更に含む処理を前記コンピュータに実行させる事を特徴とする請求項４に記載のプログラム。
【請求項７】　ネットワーク上の文書間の関係を判定する制御をコンピュータに実行させるプログラムであって、
第１の文書からリンク関係を抽出し、
前記リンク関係に基づいて、前記第１の文書からリンクされる第２の文書が、前記第１の文書の内容に関連する関連非テキスト文書であるか否か判定する、
ことを含む処理を前記コンピュータに実行させる事を特徴とするプログラム。
【請求項８】　ネットワーク上の文書が提供するサービスの種別を判定する制御をコンピュータに実行させるプログラムであって、
前記文書からユーザ入力を指定するタグを抽出し、
前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定する、
ことを含む処理を前記コンピュータに実行させる事を特徴とするプログラム。
【請求項９】　ネットワーク上から文書を検索する文書検索方法であって、
前記ネットワークから文書を収集し、
前記文書からリンク関係を抽出し、
第１の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、
前記抽出された各文書の人気度を算出し、
検索条件に基づいて文書を検索し、
前記検索された文書を前記人気度に基づいてランキングし、
前記ランキング結果に基づいて、前記検索された文書に関する情報を出力する、
ことを含むことを特徴とする文書検索方法。
【請求項１０】　第２の期間内に算出された前記人気度に基づいて、前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出し、
前記人気変化度に関する情報を前記検索された文書に関連する情報に加える、
ことを更に含むことを特徴とする請求項９に記載の文書検索方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ネットワーク上に存在する文書の整理に関し、特に文字情報のみならず、画像、音声等の様々な形態の大量の文書が存在し、かつそれらの文書が激しく変化するような場合に好適な文書整理技術に関する。
【０００２】
【従来の技術】例えば、WWW（WorldWide Web、以下、ウェブという）は、急成長しているインターネットリソースである。ウェブには、２０００年において20億ページ以上という調査があるように、大量の文書（ウェブページともいう）が存在する。また、ウェブには、存在する文書が大量であるだけでなく、非常に文書の変化が速いという特徴もある。
【０００３】ウェブ・アーカイブ・オーガニゼーション（Web Archive Organization）による調査では、ウェブでは、情報が毎月10%ずつ増加し、一つの文書の寿命（文書が作成されてからメンテナンスされなくなるまで）は約７５日という結果もある。
【０００４】現在、このようなウェブ上に存在する情報を検索する検索サービスが、いくつか提供されている。この検索サービスにおいて、検索の結果得られた文書のネットワーク上の位置を示す情報、例えば、ＵＲＩ（Uniform Resource Identifier）又はＵＲＬ（Uniform Resource Locator）、とそのウェブページの内容を説明する文が、検索者に提供される。
【０００５】また、近年、ブロードバンド時代を反映し、文書のコンテンツはテキストから動画・音声等に、また単に内容を閲覧させる文書からサービスを提供する文書に、文書の内容が移行している。
【０００６】

【発明が解決しようとする課題】しかし、従来の検索サービスでは、ある時点でのウェブの状況に基づいて検索サービスを提供しているため、文書が時系列的にどのような状況にあるか、例えば人気が出始めであるのか、定番的なものであるか、人気が落ちているものであるのかは不明であるという問題があった。例えば、ウェブから「最近人気のあるウェブページ」を調べる方法はなかった。

【０００７】また、ウェブの場合、古くなった文書を作者が削除したり、文書の内容をこまめに更新したりすることはあまりない。そのため、単純に文書へリンクしている他の文書の数（被リンク数）に基づいて、文書の人気の高さの度合い、つまり人気度を算出すると、人気度が減るということは殆んどないという問題もあった。

【０００８】また、ブロードバンド時代を反映して文書が、テキスト中心から、画像などの非テキストやサービスを含んだものが中心になっているが、その変化に対応した文書の整理方法がなかった。

【０００９】以上の問題を鑑み、単純な被リンク数に基づく文書の人気度が増加する一方で減少することがないという問題を解決することを１つの目的とする。また、文書の人気度が時系列的にどのような状況にあるのかを示す情報を得る事を可能とすることを更なる目的とする。また、文書の内容等の移行に対応して文書を整理する事を可能とすることを更なる目的とする。

【００１０】

【課題を解決するための手段】本発明の１態様によれば、ネットワーク上の文書の人気の高さの度合いである人気度を算出する人気度算出方法において、文書からリンク関係を抽出し、第１の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出することを特徴とする。

【００１１】第１の期間内に収集された文書又は第１の期間内に更新された文書を対象として人気度を算出することにより、古い文書を人気度を算出する対象から省き、ひいては、文書の人気度が増加する一方で減少することがないという問題を解決する。第１の期間は、有意な人気度を算出するために、ある程度長い期間、例えば１５０日程度であることが望ましい。

【００１２】ここで、前記リンク関係及び前記文書の前記ネットワーク上の位置を示す文書位置情報に基づいて前記人気度を算出することしてもよい。これにより、文書の内容を見ることが不要であるため、人気度を迅速に算出することが可能となる。

【００１３】上記方法において、更に、第２の期間内に算出された前記人気度に基づいて、前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出することとしてもよい。これにより、文書の人気度が時系列的にどのような状況にあるのかを示す情報を得る事が可能となる。

【００１４】ここで、第２の期間は、人気度の変化を見るために、あまり長い期間でない、例えば数週間程度である事が望ましい。上記方法において、前記第２の期間内に算出された前記人気度の時間に対する回帰式を算出し、前記人気変化度を前記回帰式に基づいて算出することとしてもよい。この場合、前記回帰式の回帰係数に基づいて前記人気変化度を決定することとしてもよいし、前記回帰式の切片に基づいて、前記人気度の時間に対する推移の傾向を決定することとしてもよい。

【００１５】また、回帰式を算出する際に、人気度の代わりに、前記抽出された文書の人気度に基づく順位を用いることとしてもよい。また、本発明の別の１態様によれば、ネットワーク上の文書間の関係を判定する文書関係判定方法において、第１の文書からリンク関係を抽出し、前記リンク関係に基づいて、前記第１の文書からリンクされる第２の文書が、前記第１の文書の内容に関連する非テキスト文書であるか否か判定することを特徴とする。これにより、近年多くなっている、画像など非テキストメディアの種別に応じて、文書を整理することが可能となる。

【００１６】上記方法において、前記第１の文書から前記第２の文書にリンクする部分の近辺にある文字列を前記第１の文書から抽出し、前記文字列に基づいて、前記第２の文書が前記第１の文書の内容に関連する関連非テキスト文書であるか否か判定することを更に含むこととしても良い。例えば、文字列が、ＭＥＰＧ、動画、ストリーミング等、第２の文書が非テキストフォーマットであることを示す文字列である場合、第２の文書は第１の文書の内容に関連する非テキスト文書であると推定できる。

【００１７】また、上記方法において、前記拡張子が特定の拡張子でない場合、前記第２の文書が前記第１の文書の内容に関連する非テキスト文書でないと決定することを含むこととしてもよい。拡張子は、第２の文書の文書フォーマットを示すため、これに基づいて非テキスト文書であるか否か判定する事ができる。

【００１８】また、上記方法において、前記第２の文書が前記第１の文書内で所定回数以上使用されているか否かに基づいて、前記第２の文書は前記第１の文書の内容に関連する非テキスト文書であるか否か判定することとしてもよい。例えば、ブリット等は、画像であるが、これらの文書作成用の素材系の画像は、１つの文書中で何度も繰り返して使用されることが多いため、第１の文書中での使用回数が多い第２の文書は第１の文書の内容に関連していないと推定する事が可能である。

【００１９】また、上記方法において、前記第１の文書内に前記第２の文書のファイル名と類似したファイル名を持つ第３の文書がある場合、前記第２の文書の前記ファイル名が前記第３の文書の前記ファイル名よりも辞書順に若くない場合、前記第２の文書を第１の文書の内容

に関連する非テキスト文書としてデータベースに登録しないことを更に含むこととしてもよい。

【００２０】例えば、第１の文書が写真集である場合、多くの画像を含む。これらの画像を全て第１の文書の内容に関連する非テキスト文書として登録すると、かえって煩雑となる可能性がある。しかし、この場合、画像ファイルのファイル名が互いに類似している事が多いため、複数の文書のファイル名のうち最も辞書順にファイル名が若い文書のみを第１の文書の内容に関連する非テキスト文書として登録することにより、このような煩雑さを解消する事ができる。

【００２１】また、上記方法において、前記第２の文書からリンクされる第３の文書がある場合、前記第１の文書の前記ネットワーク上の位置を示す文書位置情報と前記第２の文書の文書位置情報に基づいて、前記第２の文書が前記第１の文書の内容に関連する非テキスト文書であるか否か判定することを更に含むこととしても良い。また、前記第１の文書の前記文書位置情報と前記第３の文書の文書位置情報に基づいて、前記第２の文書が前記第１の文書の内容に関連する非テキスト文書であるか否か判定することを更に含むこととしてもよい。

【００２２】例えば、第１の文書中には、バナー広告等、文書の内容に関係ない非テキスト文書が第２の文書として含まれることがある。このような場合、前記第２の文書の前記文書位置情報と第２の文書のリンク先である第３の文書の前記文書位置情報が、前記第１の文書の前記文書位置情報と同じサーバアドレス又はドメインを持たないことが多いため、各文書の文書位置情報に基づいて、広告バナーのような第１の文書の内容に関連しない非テキスト文書を除く事ができる。

【００２３】また、本発明の更なる別の１態様によれば、ネットワーク上の文書が提供するサービスの種別を判定するサービス種別判定方法において前記文書から、ユーザ入力を指定するタグを抽出し、前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定することを含むことを特徴とする。これによっても、近年の文書の内容等の変化に対応して、より具体的には、文書が提供するサービスの種別に応じて文書を整理することが可能となる。ユーザ入力を指定するタグとして、例えば、文書を記述する言語がＨＴＭＬである場合、フォームタグが挙げられる。

【００２４】上記方法において、前記文書にユーザ入力を指定するタグが含まれていない場合、前記文書はサービスを提供しないと決定することを更に含むこととしてもよい。文書中に何もユーザ入力欄が含まれていない場合、その文書がサービスを提供している可能性は低いからである。

【００２５】また、文書に含まれるボタンの表示に基づいて、前記文書が提供するサービスの種別を判定することを更に含むこととしてもよい。また、さらに、ボタンの表示に加えて入力欄に基づいて、前記文書が提供するサービスの種別を判定することを更に含むこととしてもよい。文書が提供するサービスによって、多くの場合、ボタン等の入力欄の形式が決まっているからである。

【００２６】より具体的には、例えば、前記文書に、商品を購入する旨を示す表示をもつボタンが含まれている場合、前記文書が提供するサービスの種別を販売店として決定することを更に含むこととしてもよい。商品を販売するサービスを提供する文書において、商品の注文を受けるために、このようなボタンが含まれる事が多いからである。

【００２７】また、例えば、前記文書に、ユーザ入力エリア、及び検索を示す表示をもつボタンが含まれている場合、前記文書が提供するサービスの種別を検索として決定することを更に含むこととしてもよい。

【００２８】また、本発明の各態様にかかわる方法において行われる手順を実現する手段を備える装置によっても、前述した方法と同様の作用・効果を得ることが可能である。また、上述した本発明の各方法において行なわれる手順と同様の制御をコンピュータに行なわせるプログラムをコンピュータに実行させる事によっても、前述した方法と同様の作用・効果を得ることが可能である。また、上述のプログラムを記録したコンピュータ読み取り可能な記録媒体から、そのプログラムをコンピュータに読み出させて実行させることによっても、前述した方法と同様の作用・効果を得ることが可能である。

【００２９】

【発明の実施の形態】以下、本発明の実施の形態を図面に基づいて説明する。図１に、本発明の原理を示す。本発明に係わる文書整理装置は、リンク関係に基づいて、文書の人気の高さの度合いを示す人気度を算出し、さらに、その人気度が時系列的にどのように変化しているのかを示す人気変化度を算出する。そして、各文書を算出された人気度及び人気変化度に基づいて整理する。

【００３０】図１に示すように、文書整理装置１０は、人気度算出手段１１、人気度遷移算出手段１２を備える。人気度算出手段１１は、第１の期間に収集されたネットワーク上の文書間のリンク関係に基づいて、各文書の人気の高さの度合いを示す人気度を算出する。ここで、人気度算出手段１１は、第１の期間内に収集された文書又は第１の期間内に更新された文書を対象として人気度を算出する。これにより、文書の人気度が増加する一方で減少することがないという問題を解決する。

【００３１】人気度遷移算出手段１２は、第２の期間内に人気度算出手段１１によって算出された人気度に基づいて、人気度の変化の方向と度合いを示す人気変化度を算出する。なお、人気度遷移算出部１２は、人気変化度を算出する際に、人気度の代わりに、人気度に基づいて各文書をランキングした人気度順位を用いることとしても良い。これにより、ネットワーク上の文書の人気度が

時系列的にどのように変化しているのか解析する事が可能となる。

【００３２】また、近年、ブロードバンドインターネット時代を反映して、文書の内容（コンテンツ）はテキストから画像、動画、音声等のような非テキストに、さらに、単に情報を読ませる文書から検索や登録などサービスを提供する文書に重点が移りつつある。しかし、例えば、従来の検索サービスにおいて、検索結果として、文書のネットワーク上の位置を示す情報とその文書の内容を説明する文とを検索者に提供するだけでは、その文書がどのような非テキストコンテンツを含んでいるのか、あるいは、その文書でどのようなサービスを行っているかは、その文書にアクセスしない限り、その検索者にはわからない。

【００３３】また、このような非テキストコンテンツを整理する際に、単純にファイルの拡張子に基づいて文書に含まれる非テキストコンテンツを判定すると、その文書に含まれているバナーやブリット(点)など、その文書の内容とは関連のない非テキストコンテンツもその文書と関連するコンテンツとして整理されてしまうという問題もある。

【００３４】そこで、図１に示すように、本発明に係わる文書整理装置１０は、更に、関連非テキストコンテンツ判定手段１３及びサービス種別判定手段１４とを備える。関連非テキストコンテンツ判定手段１３は、文書間のリンク関係に基づいて、各文書に含まれる非テキストコンテンツのうち、その文書の内容に関連する非テキストコンテンツを判定し、文書の内容に関連すると判定された非テキストコンテンツをその文書に対応させて整理する。

【００３５】サービス種別判定手段１４は、各文書に含まれるタグ、例えば入力欄を作成する際に用いるユーザ入力を指定するタグ、例えば、ＨＴＭＬの場合のフォームタグ等に基づいて、その文書がサービスを提供しているか否か判定し、更に文書がサービスを提供している場合そのサービスの種別を判定し、判定したサービス種別をその文書に関連させて整理する。これにより、例えば、検索サービスにおいて、検索結果として、文書のネットワーク上の位置を示す情報とその文書の内容を説明する文に加えて、その文書の内容と関連する非テキストコンテンツ及びその文書で提供されているサービスについての情報を、その文書に関する情報として提供する事等が可能となる。

【００３６】以下、本発明の実施形態について説明する。なお、上述の文書整理装置をネットワーク上から文書を検索する文書検索装置に適用した場合について説明するが、本発明の適用範囲を限定する趣旨ではない。

【００３７】図２に、本発明の実施形態に係わる文書検索装置の構成を示す。文書検索装置１００は、ネットワークから文書を収集し、収集された文書を整理する。ネットワークとして、イントラネットや専用回線等のＬＡＮ（Local Area Network）、公衆回線やインターネット等のＷＡＮ（Wide Area Network）が考えられる。文書検索装置１００は、直接又は、不図示のネットワークを介して接続された端末（不図示）のユーザからの指示に従って、文書を検索し、検索結果をユーザに提供する。

【００３８】なお、文書検索装置１００がネットワークを介して端末にサービスやデータを提供するサーバである場合、ユーザの端末はブラウザ１０８を備え、ユーザは、ブラウザ１０８を用いて文書検索装置１００から送信される情報を閲覧することとしてもよい。

【００３９】図２に示すように、文書検索装置１００は、収集部１０１、人気度算出部１０２、人気度遷移算出部１０３、関連非テキストコンテンツ判定部１０４、サービス種別判定部１０５、ページ分類部１０６、検索サービス部１０７、文書テーブル１１、リンク関係テーブル１１２、人気度テーブル１１３、人気度変化テーブル１１４、非テキストコンテンツテーブル１１５及びサービス種別テーブル１１６を備える。収集部１０１、人気度算出部１０２、人気度遷移算出部１０３、関連非テキストコンテンツ判定部１０４、サービス種別判定部１０５、ページ分類部１０６及び検索サービス部１０７は、例えば、プログラムにより記述されたソフトウェアコンポーネントに対応し、文書検索装置１００を実現するコンピュータのメモリの特定のプログラムコードセグメントに格納される。

【００４０】ここで、ネットワーク上に存在する文書、つまりウェブページを記述する言語として、例えば、ＨＴＭＬ（HyperText Markup Language）、ＸＨＴＭＬ（eXtensible HyperText MarkupLanguage）、ＸＭＬ（eXtensible HyperText Markup Language）、ＳＧＭＬ（Standard Generalized Markup Language）等のような、リンク関係を文書内に埋め込む事が可能な言語が考えられる。また、本発明では、上記のような言語で記述されたテキスト文書以外に、画像、動画、音声等も文書として扱う。以下、テキスト文書を記述する言語をHTMLであると仮定して説明する場合もあるが、本発明を限定する趣旨ではない。

【００４１】収集部１０１は、ネットワーク上で公開されている文書を収集し、収集された文書に、文書を識別する文書ＩＤ（IDentification information）を付す。さらに、収集部１０１は、収集された文書のリンク関係を解析する。そして、収集部１０１は、収集された文書のネットワーク上の位置を示す文書位置情報を文書テーブル１１１に格納し、収集された文書間のリンク関係に関する情報をリンク関係テーブル１１２に格納する。

【００４２】ここで、文書位置情報として、例えば、ＵＲＩ（Uniform Resource Identifier）等が考えられる。なお、ＵＲＩは包括的な概念であり、現在は、ＵＲＩの機能の一部を具体的に使用したＵＲＬ（UniformRes

ource Locator）が広く利用されている。以下、文書位置情報をＵＲＬであると仮定して説明する場合もあるが、本発明を限定する趣旨ではない。

【００４３】人気度算出部１０２は、定期的（又は不定期的に）に、収集部１０１によって収集された文書のリンク関係に基づいて、人気の高さの度合いを示す人気度を算出し、算出結果を人気度テーブル１１３に格納する。人気度を算出する際に、人気度算出部１０２は、収集部１０１によって収集された文書のうち、第１の期間内に収集された文書又は第１の期間内に更新された文書を、人気度を算出する対象となる文書とする。ここで、第１の期間は、あまり短期間では人気度として意味のある結果を得る事ができないため、ある程度長い期間である必要がある。例えば、第１の期間として、人気度を算出する日の前１５０日間が考えられる。

【００４４】これにより、作成された後、更新されずに放置されたままの文書を人気度を算出する対象から省く事が可能となる。延いては、ある文書の人気度を単純に時系列に算出すると、人気度が単調に増加する一方であるという問題を解決する事ができる。

【００４５】人気度遷移算出部１０３は、第２の期間内に人気度算出部１０２が算出した人気度に基づいて、各文書について人気度の変化の方向と度合いを示す人気変化度を算出し、算出結果を人気度変化テーブル１１４に格納する。ここで、第２の期間は、あまり長いと、短期的な人気度の変動を把握する事ができないため、ある程度短い期間、例えば、数週間程度である必要がある。例えば、第２の期間として、人気変化度を算出する日の前１４日間が考えられる。

【００４６】より具体的には、例えば、人気度遷移算出部１０３は、各文書について、第２の期間内に算出された人気度を、人気度テーブル１１３から取得し、取得された人気度の時間に対する線形回帰式を算出し、その線形回帰式の回帰係数を人気変化度として得る。また、人気度遷移算出部１０３は、人気変化度を算出する際に、人気度の代わりに、人気度に基づいて各文書をランキングした人気度順位を用いることとしても良い。これにより、ネットワーク上の文書の人気度が時系列的にどのように変化しているのか解析する事が可能となる。

【００４７】関連非テキストコンテンツ判定部１０４は、各文書の文書位置情報に含まれるファイル名の拡張子や、文書中のリンクが埋め込まれた部分の前後にある文字列に基づいて、各文書のタイプを判定する。更に、関連非テキストコンテンツ判定部１０４は、文書間のリンク関係に基づいて、各文書に含まれる非テキストコンテンツが、各文書の内容に関連するか否か判定する。そして、関連非テキストコンテンツ判定部１０４は、各文書の内容に関連すると判定された非テキストコンテンツを、その文書に対応させて非テキストコンテンツテーブル１１５に格納する。これにより、各文書に含まれる非テキストコンテンツのうち、その文書の内容に関連しない非テキストコンテンツを除去し、その文書の内容に関連する非テキストコンテンツを文書に対応させて整理することが可能となる。

【００４８】サービス種別判定部１０５は、各テキスト文書に含まれる入力欄を記述する情報に基づいてその文書で提供するサービスの種別を判定し、判定されたサービス種別をその文書に対応させてサービス種別テーブル１１６に格納する。これにより、各文書が提供するサービスの種別を文書に対応させて整理することが可能となる。

【００４９】ページ分類部１０６は、関連分野等に基づいて各文書を分類する。文書の分類方法については、既に様々な分類技術が存在するため、本実施形態では詳しく説明することを省略する。

【００５０】検索サービス部１０７は、ユーザの指示に従って、ネットワーク上の文書を検索し、検索結果をユーザに提供する。その際に、検索サービス部１０７は、検索の結果得られた文書に関する情報を、人気度テーブル１１３及び人気度変化テーブル１１４から取得し、検索された文書の内容を説明する情報及び文書位置情報に加えて、人気度、人気変化度をユーザに提供する。これにより、ユーザは、検索された文書の人気が、今どのような状態にあるのか、人気が出始めであるのか、人気が落ちてきているのか、検索結果の出力画面で提供される情報によって知ることができる。

【００５１】さらに、検索サービス部１０７は、検索の結果得られた文書に関する情報を、非テキストコンテンツテーブル１１５及びサービス種別テーブル１１６から取得し、検索された文書の内容に関連する非テキストコンテンツに関する情報及び検索された文書で提供されているサービス種別に関する情報もユーザに提供することとしてもよい。これにより、ユーザは、検索の結果得られた文書が、どのような非テキストコンテンツを含むのか、或いは、その文書でどのようなサービスが提供されているのか、その文書にアクセス（閲覧する）しなくとも、検索結果の出力画面で提供される情報によって知ることが可能となる。

【００５２】また、ユーザが、１以上の文書の人気度に関する情報を提供するよう要求した場合、検索サービス部１０７は、その文書に関する情報を人気度テーブル１１３、人気度変化テーブル１１４等から取得し、取得された情報を時系列に提供することとしてもよい。これにより、ユーザは、ある文書の人気度の推移を分析することが可能となる。

【００５３】以下、図３から図８を用いて、各テーブルのデータ構造について説明する。まず、図３を用いて文書テーブル１１１のデータ構造について説明する。図３に示すように、文書テーブル１１１は、各文書について文書位置情報とそれに対応する文書ＩＤを格納する。こ

れにより、各文書の文書位置情報は文書ＩＤに変換され、以降の処理では文書ＩＤを用いて各文書のリンク関係等に関する情報を管理することが可能となる。

【００５４】次に、図４を用いて、リンク関係テーブル１１２のデータ構造について説明する。リンク関係テーブル１１２は、各文書についてのリンク関係情報を格納する。図４に示すように、リンク関係情報は、その文書が収集された日時（又は日付）、更新された日時（又は日付）、リンク元となっている文書の文書ＩＤ、リンク先となっている文書の文書ＩＤを項目として含む。以下の説明において、リンク元となっている文書の文書ＩＤをリンク元ＩＤといい、リンク先となっている文書の文書ＩＤをリンク先ＩＤということとする。なお、各文書の更新日時が取得困難な場合、収集日時を更新日時に代えて扱う事としてもよい。

【００５５】次に、図５を用いて、人気度テーブル１１３のデータ構造について説明する。人気度テーブル１１３は、各文書についての人気度情報を格納する。図５に示すように、人気度情報は、人気度が算出された日時（又は日付）、その文書の文書ＩＤ、算出された人気度及び、人気度に基づいて文書をソートした結果である人気度順位を項目として含む。

【００５６】次に、図６を用いて、人気度変化テーブル１１４のデータ構造について説明する。人気度変化テーブル１１４は、各文書について人気度変化情報を格納する。人気度変化情報は、その文書の文書ＩＤ、人気度について線形回帰式を算出した結果得られた回帰係数（傾き）及び切片、並びに、人気度順位について線形回帰式を算出した結果得られた回帰係数（傾き）及び切片を、項目として含む。

【００５７】次に、図７を用いて、非テキストコンテンツテーブル１１５のデータ構造について説明する。非テキストコンテンツテーブル１１５は、リンク先を持つ文書について、その文書の文書ＩＤと、その文書の内容に関連し、その文書からリンクされている非テキストコンテンツの文書ＩＤ（以下、関連非テキストコンテンツＩＤという）と、その非テキストコンテンツのファイル種別を格納する。

【００５８】最後に、図８を用いて、サービス種別テーブル１１６のデータ構造について説明する。図８に示すように、サービス種別テーブル１１６は、各文書について文文書ＩＤと、その各文書で提供するサービスの種別を格納する。

【００５９】以下、図９から図１５を用いて、文書検索装置１００を構成する各部によって行われる処理について説明する。なお、ページ分類部１０６によって行われる処理についての説明は、上述のように省略する。

【００６０】まず、収集部１０１は、継続してネットワークから文書を収集し、収集された文書間のリンク関係を解析し、収集及び解析結果を文書テーブル１１１及びリンク関係テーブル１１２に格納する。人気度算出部１０２は、定期的に、例えば毎日、算出日の前の一定期間内に収集又は更新された文書について、人気度を算出する。なお、１日毎は例示に過ぎず、本発明を限定する趣旨ではない。以下、図９を用いて人気度を算出する処理の手順について説明する。

【００６１】図９に示すように、まず、人気度算出部１０２は、毎日、定時に起動する。人気度を算出する人気度算出日をｄ１とすると、人気度算出部１０２は、ｄ１からＮ日前、例えば１５０日前の日ｄ２を算出対象開始日として決定する（ステップＳ１１）。なお、１５０日は例示に過ぎない。Ｎは、人気度として意味のある結果を得ることができる程度に長い期間であればよい。

【００６２】続いて、人気度算出部１０２は、収集日又は更新日が算出対象開始日ｄ２から算出日ｄ１までの間にあるリンク関係情報をリンク関係テーブル１１２から抽出する（ステップＳ１２）。人気度を算出する対象となる文書の収集日又は更新日を一定期間内に制限することにより、作成された後、更新されずに放置されたままの文書を人気度を算出する対象から除く事が可能となる。

【００６３】人気度算出部１０２は、抽出したリンク関係情報のうちで、同じリンク元ＩＤを持つリンク関係情報がある場合、最新の収集日又は更新日を持つリンク関係情報を残し、その他の同じリンク元ＩＤを持つリンク関係情報を削除する（ステップＳ１３）。これにより、同じ文書について人気度を重複して算出することを防ぐ事が可能となる。

【００６４】人気度算出部１０２は、抽出したリンク関係情報に基づいて、各文書の人気度を算出する（ステップＳ１４）。より具体的には、人気度算出部１０２は、文書の内容を参照することなく、リンク関係及び、リンク元の文書とリンク先の文書の文書位置情報を示す文字列の類似している度合いである類似度に基づいて、各文書の人気度を算出する。以下、人気度の算出手順について説明する。

【００６５】人気度を算出する際の基本的な考え方は以下の通りである。

１．類似していない文書位置情報を持つ文書から多くリンクされている文書は、人気が高い。

【００６６】例えば、一般に、同一サイト内に設けられた複数の文書はそのサイト内の他の文書にリンクされているが、それらの文書の文書位置情報は相互に類似する。従って、文書位置情報を示す文字列が相互に類似している文書からリンクされている文書の人気度は低いと推定できるからである。

【００６７】２．多くの文書からリンクされている文書ほど人気度が高い文書であり、類似していない文書位置情報を持ち人気度が高い文書からリンクされている文書の人気度は高い。

【００６８】例えば、有名なディレクトリサービス等及び官公庁等は、多くの文書からリンクされているが、このような文書からリンクされている文書の方が、個人が開設するサイトやそのコンテンツのエントリページからリンクされている文書よりも人気度が高いと考えられるからである。また、多くの文書やミラーサイトを抱えるサービス（サイト）に設けられた文書等はそのサイト内でリンクされていることが多い。１つのサイト内の文書の文書位置情報は、例えばドメインが同じ等大抵類似しているため、「文書位置情報が類似していない文書からリンクされている文書の人気度は高い」という考え方を導入すれば、サイト内で多数回リンクしあっている文書の人気度が高くなってしまうことを解消することが可能となる。

【００６９】３．文書位置情報が類似しているか否かは、サーバアドレス、パス、ファイル名の全てが異なるものが最も小さく、ミラーサイトや同一サーバ内の文書は類似度が高くなるように、文書位置情報を示す文字列から定義する。

【００７０】上述の３つの考え方を導入することにより、全てのリンク関係を同等に扱わないで、リンク関係に重みを与えて扱うこととしている。より具体的には、リンク関係に重みをリンク元とリンク先文書の文書位置情報の類似度の逆数として与えることとしている。

【００７１】以下、人気度を算出する手順についてより詳しく説明する。

人気度の算出対象となる文書集合をＤＯＣ＝｛p1，p2,....pN｝、

文書ｐの人気度をWp 、文書ｐのリンク先の文書集合をRef(p)、文書ｐのリンク元の文書集合をRefed(p)、文書ｐと文書ｑの文書位置情報の類似度をsim(p,q)、相異度をdiff(p,q)＝1/sim(p,q)とすると、文書ｐから文書ｑにリンクが張られているとした時、そのリンク関係の重みlw(p,q)を以下の（１）式で定義する。

【００７２】

【数１】

$$lw(p,q)=\frac{diff(p,q)}{\sum_{i\in Ref(p)} diff(p,i)}=\frac{1}{sim(p,q)\sum_{i\in Ref(p)}\frac{1}{sim(p,i)}} \quad \cdots (1)$$

【００７３】この（１）式から分かるように、lw(p,q)は、文書ｐと文書ｑのＵＲＬの類似度sim(p,q)が低いほど、また、文書ｐから文書ｐへののリンクの数がより少ないほど大きくなる。

【００７４】文書ｑの人気度Wq は、各文書ｐ∈ＤＯＣに対して、Cq を定数（人気度の下限であり、文書によって異なる値を与えてもよい。）として、

【００７５】

【数２】

$$W_q=C_q+\sum_{p\in Refed(q)} W_p\times lw(p,q) \quad \cdots (2)$$

【００７６】という（２）式に示す連立一次方程式の解として定義される。人気度算出部１０２は、この連立一次方程式を解くことにより、各文書の人気度を算出する。なお、このような連立一次方程式の解法については、既存のアルゴリズムが多数存在するため、説明は省略する。（１）式中の文書位置情報の類似度sim(p,q)の算出方法については後述する。（１）式及び（２）式から、上述の考え方が実現されていることを読み取ることができる。すなわち、（１）式から文書位置情報の類似度が低ければ、リンク関係の重みｌｗは大となる。そして、（２）式からリンク関係の重みｌｗが大きい文書からリンクされている文書の人気度Wqは、高くなる。つまり、類似度の低い文書位置情報を持つ文書から多くリンクされている文書の人気度は、高くなる。また、（２）式から多くの文書からリンクされている文書ほど人気度が高くなる。さらに、（２）式から人気度Wが高い文書からリンクされている文書の人気度は高くなることも分かる。

【００７７】次に、（１）式及び（２）式中の文書ｐと文書ｑの文書位置情報の類似度sim(p,q)について説明する。以下、文書位置情報をＵＲＬと仮定して説明するが、本発明を限定する趣旨ではない。

【００７８】一般に、文書のＵＲＬは、サーバアドレス、パス、ファイル名の三種類の情報から構成される。例えば、ＷＷＷ文書のＵＲＬ、http://www.flab.fujitsu.co.jp/hypertext/news/1999/product1.html は、サーバアドレス（www.flab.fujitsu.co.jp）、パス（hypertext/news/1999）、ファイル名（product1.html）の３種類の情報から構成される。

【００７９】また、サーバアドレスは、さらに“．”により階層化されており、後ろに行くにしたがって、段々広くなる。例えば、サーバアドレスがwww.flab.fujitsu.co.jpであれば、後ろから、日本(jp)、会社(co)、富士通(fujitsu)、研究所(flab)、マシン(www)という階層を表している。

【００８０】本実施形態に係わるリンク関係の重みの算出方法は、以下のような考え方に基づいている。

１．往々にして、同じような文書を同一ディレクトリに入れるため、同一サーバでパスも同じ文書位置情報は内容が似ていることが多い。

２．アクセスを分散させるために設けられるミラーサイト内の文書と、オリジナルサイトの文書の文書位置情報は類似度が高い。例えば、サーバアドレス部分だけが異なり、残りのパスやファイル名は同じ場合が多い。

３．サーバアドレス、パス、ファイル名が全てことなる文書位置情報は、類似度が低い。

【００８１】本実施形態では、与えられた２つの文書ｐ

及び分書qの文書位置情報の類似度を、上述のサーバアドレス、パス、ファイル名の三種類の組合せにより定義する。類似度sim(p,q)として、例えば、以下に述べるドメイン類似度sim-domain(p,q) 及び融合類似度sim-merge(p,q)が考えられる。

【００８２】ドメイン類似度sim-domain(p,q) は、ドメインの類似に基づいて算出される。ドメインとは、サーバアドレスの後半部分であり、会社や組織を表す。サーバアドレスが.com、.edu、.org等で終わる米国サーバの

$$\text{sim-domain}(p,q) = 1 / \alpha = 1$$

ここで、αは定数で、０より大きく１より小さい実数値を取るとする。sim-domain(p,q)の概念を導入することにより、異なるドメインを持つ文書が検索されやすくなる。言い換えると、同じドメインを持つ文書は検索されにくくなる。

【００８４】sim(p,q)として、前述の三種類の情報を融合した融合類似度sim-merge(p,q)を次のように定義する。

sim-merge(p,q)＝（サーバアドレスの類似度）＋（パスの類似度）＋（ファイル名の類似度）

以下、右辺の各項の算出方法について説明する。

【００８５】サーバアドレスの類似度は、アドレスの階層を後ろから見ていき、ｎレベルまで一致した場合、類似度を１＋ｎとする。例えば、www.fujitsu.co.jp とwww.flab.fujitsu.co.jpは、相互に３レベルまで一致しているので、融合類似度は４となる。www.fujitsu.co.jpとwww.fujitsu.com は、相互に１レベルも一致していないので（一致０レベル）、融合類似度は１である。

【００８６】パスの類似度は、先頭からパスの"/" で区切られた要素毎に比較し、一致したレベルまでを類似度とする。例えば、/doc/patent/index.htmlと/doc/patent/1999/2/file.htmlとは、２レベルまで一致しているので類似度は３である。

【００８７】ファイル名の類似度は、ファイル名が一致する場合、類似度１とする。このsim-merge(p,q)によっても、ＵＲＬが似通った文書からリンクされている文書の人気度は、ＵＲＬが似通っていない文書からリンクされている場合と食らえて低くなる。従って、lw(p,q)の中にsim(p,q)又はdiff(p,q)という概念を導入することにより、大量の文書を抱えるサーバ（サイト）や個人が単に量が多いというだけで人気度が高いことになるという問題を解消することができる。

【００８８】人気度を算出した後、人気度算出部１０２は、各文書を人気度が高い順にソートする事により人気度順位を取得する（ステップＳ１５）。人気度順位の時系列の変化は、増加することもあれば、減少することもある。従って、従来の算出方法による人気度の時系列の変化は増加する一方であったという問題は、人気度の代

場合はサーバアドレスの後ろから２つめまで、サーバアドレスが.jp、.fr 等で終わる他国のサーバの場合はサーバアドレスの後ろから３つめまでがドメインに相当する。例えば、www.fujitsu.com のドメインはfujitsu.comであり、www.flab.fujitsu.co.jpのドメインはfujitsu.co.jp である。

【００８３】文書ｐと文書ｑのドメイン類似度は以下の（３）式により定義される。

（ｐ、ｑが同一ドメインの場合）
（ｐ、ｑが異なるドメインの場合）
・・・（３）

わりに人気度順位の時系列の変化に注目する事によっても解決する事が可能となる。最後に、人気度算出部１０２は、算出した人気度及び人気度順位を、各文書の文書ＩＤ及び人気度算出日とともに人気度テーブル１１３に格納し（ステップＳ１６）、処理を終了する。

【００８９】例えば、ユーザに文書を検索した結果を提供する際に、上述のように算出した人気度に基づいて各文書は、ソート又はランキングされることとしてもよい。また、ある文書に関する情報を提供する際に、その文書の人気度もユーザに提供することとしても良い（後述）。

【００９０】以下、図１０を用いて、人気度の算出における本発明の特徴について説明する。図１０（ａ）は、従来の算出方法によって算出した人気度の時間的変化を示す図である。図１０（ａ）において、横軸は時間、縦軸は人気度を示す。ウェブにおいて一度作成された文書を作者が削除したり、更新したりすることはあまりないため、従来技術のように単純に文書へリンクしている他の文書の数（被リンク数）に基づいてその文書の人気度を算出すると、人気度が減ることはなく、図１０（ａ）に示すように増加する一方となる。

【００９１】図１０（ｂ）は、本発明に係わる算出方法によって算出した人気度の時間的変化を示す図である。図１０（ｂ）においても、横軸は時間、縦軸は人気度を示す。本発明によれば、算出対象開始日から人気度算出日間での間の一定期間内に収集又は更新された文書について人気度を算出するため、従来のように、一度作成された後、長期間放置されたままの文書は人気度を算出する対象とならない。従って、例えば、長期間放置されたままの文書をリンク元とする文書の人気度は、従来よりも低く算出されることとなる。これにより、従来、人気度が増加する一方であったという問題を解決する。

【００９２】また、例えば、ウェブで公開されたばかりのサイトのトップページは、そのサイト内の文書等から多くリンクされているため、そのトップページの人気度は当初高く算出されるが、その後サイト内の文書が更新されずに放置されると、そのトップページの人気度は低下し、人気度の高い状態は一過性のものとなる。

【００９３】図１０（ｂ）に示す文書の人気度は、当初、人気度が急に上昇しているが、ある程度の時間が経過した後、人気度は減少に転じ、以後減少しつづけている。このことから、この文書の流行は一過性に終わった事がわかる。

【００９４】図１０（ｃ）は、本発明に係わる算出方法によって算出した人気度に基づく人気度順位の時間的変化を示す図である。図１０（ｃ）において、横軸は時間、縦軸は人気度順位を示す。人気度順位は、人気度を算出する対象となる文書全体から見たその文書の相対的な人気度を示す情報であるため、その性質から従来の算出方法によって人気度を算出した場合でも、増加しつづける事はあまり考えられない。従って、人気度順位の時間的変化に基づいて文書の人気度を判断することによっても、従来、人気度が増加する一方であったという問題を解決することができる。

【００９５】また、本発明に係わる算出方法によって算出した人気度に基づく人気度順位の時間的変化によれば、その文書が、人気度を算出する対象となる文書全体から見て平均的な順位の推移を示す場合、図１０（ｂ）のグラフに示すように、人気度順位は、時間が経過してもほぼ一定に推移する。また、その文書の人気度が増加している場合、人気度順位も人気度の増加にあわせて上昇する。他方、その文書の人気度が減少している場合、人気度順位は、人気度の減少にあわせて下降する。一般に、文書の人気は、当初増加期から始まり、安定期を経て減少期に至る。この場合、図１０（ｃ）に示すように、人気度順位は、増加期には上昇し、安定期にいたるとほぼ一定になり、減少期には下降するため、人気度順位の時間的変化は山形になる。

【００９６】次に、図１１を用いて、人気変化度を算出する処理の手順について説明する。人気算出部１０２が人気度を算出すると、人気度遷移算出部１０３は、一定期間内に算出された人気度を人気度テーブル１１３から取得し、人気度の時間的変化である人気変化度を算出する。

【００９７】まず、人気度遷移算出部１０３は、人気度算出日ｄ１からM日、例えば１４日前の日ｄ３を算出対象開始日として決定する（ステップＳ２１）。なお、１４日は例示に過ぎない。Mは、あまり長く取ると、短期的な人気度の変動を把握する事ができなくなるため、数週間程度にすることが望ましい。

【００９８】続いて、人気度遷移算出部１０３は、各文書について算出対象開始日ｄ３から人気度算出日ｄ１の間に算出された人気度又は人気度順位を人気度テーブル１１３から取得する（ステップＳ２２）。人気度遷移算出部１０３は、各文書ごとに、人気度又は人気度順位の時間に対する線形回帰式を算出し、その線形回帰式の回帰係数及び切片ｂを得る（ステップＳ２３）。人気度に基づいて線形回帰式を算出した場合、回帰係数ａが人気変化度に相当し、人気度順位に基づいて線形回帰式を算出した場合、回帰係数ａを切片ｂで除算した値、ａ／ｂが人気変化度に相当する。

【００９９】以下、線形回帰式の算出方法について詳しく説明する。日付ｄ３からｄ１までの（ｄ３，ｄ３＋１，・・・，ｄ１）のそれぞれの日付における人気度又は人気度順位の値を（w_0，w_1，・・・w_{M-1}）とすると、線形回帰式

ｒ＝ａ（ｄ１−ｄ３）＋ｂ

は、最小二乗法によって算出される。ここで、ａは回帰係数であり、以下の式で算出される。

【０１００】

ａ＝（Ｍ×Ｉｗ−Ｉ×Ｗ）／（Ｍ×Ｉ２−$Ｉ^2$）

また、ｂは切片であり、以下の式で算出される。

ｂ＝（Ｉ×Ｉｗ−Ｗ×Ｉ２）／（$Ｉ^2$ −Ｍ×Ｉ２）

ここで、Ｉｗ、Ｗ、Ｉ及びＩ２は、それぞれ以下の式で算出される。

【０１０１】

【数３】

$$Iw=\sum_{i=0}^{M-1} i*w_i$$

【０１０２】

【数４】

$$W=\sum_{i=0}^{M-1} w_i$$

【０１０３】

【数５】

$$I=\sum_{i=0}^{M-1} i=\frac{M(M-1)}{2}$$

【０１０４】

【数６】

$$I2=\sum_{i=0}^{M-1} i^3=\frac{M(M-1)(2M-1)}{6}$$

【０１０５】最後に、人気度遷移算出部１０３は、算出した各文書の回帰係数ａ及び切片ｂを、文書ＩＤとともに人気度変化テーブル１１４に格納し（ステップＳ２４）、処理を終了する。

【０１０６】人気度に基づいて線形回帰式を算出した場合、線形回帰式の回帰係数ａが正であれば人気度が上昇中であり、その絶対値が大きいほど、人気度が上昇する速度が速いことを示す。また、切片ｂが比較的高い値以上である場合、人気度が高い水準で安定している事を示し、切片ｂが比較的低い値以下である場合、人気度が低い水準で安定していることを示す。

【０１０７】一方、人気度順位に基づいて線形回帰式を算出した場合、回帰係数ａが負であれば人気度が上昇中であり、その絶対値が大きいほど、人気度が上昇する速度が速いことを示す。また、切片ｂが比較的低い値以下である場合、人気度が高い水準で安定している事を示し、切片ｂが比較的高い値以上である場合、人気度が低

い水準で安定していることを示す。

【0108】各文書の人気変化度は、その文書についての情報をユーザに提供する際に、その文書の文書位置情報、タイトル及び内容を示す情報とともに、ユーザに提供される。提供される際、人気変化度は、数値としてではなく、人気度の変化の方向及び度合いを図示するアイコンを用いて提供される事としても良い（後述）。

【0109】次に、図12を用いて各文書の内容に関連する関連非テキストコンテンツを判定する処理について説明する。文書中には、テキストコンテンツ以外にも、画像、音声等の非テキストコンテンツが含まれる事が多い。そして、文書中に含まれる非テキストコンテンツの中には、バナー広告等、文書の内容に関係ない非テキストコンテンツもある。関連非テキストコンテンツ判定部104は、リンク関係に基づいて、文書中に含まれる非テキストコンテンツが文書の内容に関連するか否か判定する。

【0110】そのために、まず、非テキストコンテンツ判定部104は、リンク関係テーブル112を参照し、リンク先IDが格納されているリンク関係情報を抽出する。なお、抽出されたリンク関係情報のうち、同じリンク元IDを持つリンク関係情報ある場合、最新の収集日又は更新日を持つリンク関係情報のみを採用し、その他は削除する。同じ文書について同じ処理を行う事を防ぐためである。

【0111】以後、抽出されたリンク関係情報に含まれるリンク元IDによって特定されるリンク元文書Sからなる文書集合をリンク元文書集合とする。抽出されたリンク関係情報に含まれるリンク先IDによって特定される文書（つまり、リンク先文書）は、判定対象文書Cという。

【0112】ステップS31からステップS40までの手順は、各リンク元文書Sに含まれる判定対象文書Cそれぞれについて行う。まず、非テキストコンテンツ判定部104は、各リンク元文書Sから判定対象文書Cへリンクする部分の近辺に存在するリンク文字列Aを抽出する（ステップS31）。

【0113】例えば、HTMLを用いた文書の場合、非テキストコンテンツ判定部104は、アンカータグ（<a>）の前後100バイトをリンク文字列Aとして抽出することとしても良い。続いて、関連非テキストコンテンツ判定部104は、そのリンク文字列Aが特定の文字列であるか否か判定する（ステップS32）。

【0114】特定の文字列とは、例えば、「MPEG」、「動画」、「ストリーミング」、「video」、「audio」及び「mp3」や、動画等のフォーマット名など、非テキストフォーマットを示す文字列である。これらの特定の文字列を定義するテーブルは、予め、文書検索装置100に備えられているものとする（不図示）。

【0115】関連非テキストコンテンツ判定部104は、そのリンク文字列Aが特定の文字列であると判定した場合（ステップS32：Yes）、その判定対象文書Cをリンク元文書Sの内容に関連する非テキストコンテンツとして判定し、ステップS40に進む。関連非テキストコンテンツ判定部104は、その判定対象文書Cの種別及びリンク元文書SのIDとともに、その判定対象文書CのIDを関連非テキストコンテンツIDとして、非テキストコンテンツテーブル115に格納し、その判定対象文書Cについての処理を終了する。

【0116】関連非テキストコンテンツ判定部104は、そのリンク文字列Aが特定の文字列でないと判定した場合（ステップS32：No）、更に、判定対象文書Cの文書位置情報に含まれる判定対象文書Cのファイル名の拡張子が、特定の拡張子であるか否か判定する（ステップS33）。

【0117】現在のウェブでは、特定の拡張子として、例えば、以下のようなものが考えられる。なお、各拡張子についての説明は、当業者に自明であるため省略する。なお、この例示は、本発明を限定する趣旨ではない。

・音楽系のコンテンツの場合

mp3、wma、wav

・動画系のコンテンツの場合

ram、rm、rv、rmm、wmv、avi、asx、qt、mov、mpeg、mpg、fla、swf

・画像系のコンテンツの場合

jpg、jpeg

関連非テキストコンテンツ判定部104は、このような拡張子によっても、判定対象文書Cが非テキストコンテンツであるか否か判定する事ができる。これらの特定の拡張子を定義するテーブルは、予め、文書検索装置100に備えられているものとする（不図示）。関連非テキストコンテンツ判定部104は、判定対象文書Cの文書位置情報に含まれるファイル名の拡張子が特定の拡張子でないと判定した場合（ステップS33：No）、判定対象文書Cは非テキストコンテンツでないとして、その文書についての処理を終了する。

【0118】関連非テキストコンテンツ判定部104は、判定対象文書Cのファイル名の拡張子が特定の拡張子であると判定した場合（ステップS33：Yes）、更にその判定対象文書Cにリンクとして使用されているか否か判定する。この判定は、例えば、HTMLの場合タグに基づいて行う事ができる。判定対象文書Cがリンクとして使用されているとは、例えば、バナー広告画像のように、その文書を選択（クリック、或いはタッチ等）することによって他の文書を閲覧することができることを意味する。

【0119】例えば、HTMLで記述された文書中で判定対象文書C（例の場合、画像）がリンクとして使用さ

れている場合、以下のように表記されることが多い。なお、この例示は、本発明を限定する趣旨ではない。

【０１２０】<;a href="判定対象文書Ｃのリンク先文書の文書位置情報 ">;<;img src=" 判定対象文書Ｃの文書位置情報">;<;/a>;

関連非テキストコンテンツ判定部１０４は、判定対象文書Ｃ及びリンク元文書Ｓの文書ＩＤを用いて文書テーブル１１１を参照し、両者の文書位置情報を取得する。そして、関連非テキストコンテンツ判定部１０４は、判定対象文書Ｃの文書位置情報及びリンク元文書Ｓの文書位置情報に基づいて、判定対象文書Ｃが格納されているサイトが、リンク元文書Ｓが格納されているサイトと同じであるか否か判定する（ステップＳ３５）。

【０１２１】より具体的には、文書位置情報が例えばＵＲＬである場合、関連非テキスト判定部１０４は、判定対象文書ＣのＵＲＬとリンク元文書ＳのＵＲＬのサーバアドレス又はドメインに基づいて、判定対象文書Ｃが格納されているサイトとリンク元文書Ｓが格納されているサイトが同じであるか否か判定する。

【０１２２】判定対象文書Ｃが格納されているサイトとリンク元文書Ｓが格納されているサイトが同じであると判定する場合（ステップＳ３５：Ｙｅｓ）、判定対象文書Ｃは、リンク元文書Ｓの内容に関連する文書であると推測できるため、ステップＳ３７に進む（後述）。これは、判定対象文書Ｃがリンク元文書Ｓの内容と関連している場合、判定対象文書Ｃは、リンク元文書Ｓが格納されているサイトと同じサイトに格納されている事が多いからである。

【０１２３】一方、判定対象文書Ｃが格納されているサイトとリンク元文書Ｓが格納されているサイトが異なると判定する場合（ステップＳ３５：Ｎｏ）、関連非テキストコンテンツ判定部１０４は、更に、判定対象文書Ｃの文書位置情報及び判定対象文書Ｃのリンク先の文書の文書位置情報に基づいて、判定対象文書Ｃのリンク先となっている文書が格納されているサイトが、リンク元文書Ｓが格納されているサイトと同じであるか否か判定する（ステップＳ３６）。なお、判定対象文書Ｃのリンク先の文書の文書位置情報は、上記例のようにリンクを埋め込むタグ付近に記載されていることが多い。

【０１２４】判定対象文書Ｃのリンク先となっている文書が格納されているサイトが、リンク元文書Ｓが格納されているサイトと同じであると判定する場合（ステップＳ３６：Ｙｅｓ）、ステップＳ３７に進む。判定対象文書Ｃのリンク先となっている文書がリンク元文書Ｓの内容と関連していると推測されるため、判定対象文書Ｃもリンク元文書Ｓの内容と関連していると推測できるからである。

【０１２５】一方、判定対象文書Ｃのリンク先となっている文書が格納されているサイトが、リンク元文書Ｓが格納されているサイトと異なると判定した場合（ステップＳ３６：Ｎｏ）、関連非テキストコンテンツ判定部１０４は、判定対象文書Ｃは、バナー広告等、リンク元文書Ｓの内容と関連しない文書であると推定し、その判定対象文書についての処理を終了する。

【０１２６】ステップＳ３７において、関連非テキストコンテンツ判定部１０４は、判定対象文書Ｃがリンク元文書Ｓ内で所定回数、例えば、３回以上使用されているか否か判定する。なお、３回は、例示に過ぎず、本発明を限定する趣旨ではない。判定対象文書Ｃがリンク元文書Ｓ内で３回以上使用されているかと判定した場合（ステップＳ３７：Ｙｅｓ）、関連非テキストコンテンツ判定部１０４は、その判定対象文書Ｃをリンク元文書Ｓの内容に関連しないと判定し、その判定対象文書Ｃについての処理を終了する。そうでない場合、ステップＳ３８に進む。

【０１２７】例えば、判定対象文書Ｃが、リストのブリット等のフォーマット、或いは文書作成用の素材である場合、１つの文書内で複数回使用される可能性が高い。このような文書は、リンク元文書Ｓの内容とは関連がないと考えられるため、関連非テキストコンテンツとして扱わないこととする。

【０１２８】ステップＳ３７でＮｏであった場合、関連非テキストコンテンツ判定部１０４は、更に、リンク元文書Ｓのリンク関係情報に含まれるリンク先ＩＤに基づいて文書テーブル１１１からリンク元文書Ｓのリンク先文書のファイル名を取得し、リンク元文書Ｓが、判定対象文書Ｃと類似したファイル名を持つ他のリンク先文書を有するか否か判定する（ステップＳ３８）。

【０１２９】判定対象文書Ｃと類似したファイル名を持つ他のリンク先文書をリンク元文書Ｓが有しないと判定した場合（ステップＳ３８：Ｎｏ）、ステップＳ４０に進み、関連非テキストコンテンツ判定部１０４は、上述のようにして、その判定対象文書Ｃを非テキストコンテンツテーブル１１５に登録する。

【０１３０】判定対象文書Ｃと類似したファイル名を持つ他のリンク先文書をリンク元文書Ｓが有すると判定した場合（ステップＳ３８：Ｙｅｓ）、関連非テキストコンテンツ判定部１０４は、判定対象文書Ｃが、判定対象文書Ｃと、それと類似するファイル名を持つリンク先文書の中で、辞書順で最も若いファイル名を持つのか否か判定する（ステップＳ３９）。辞書順とは、例えば、アルファベットの先の順或いは、数字では小さい順ということを意味する。

【０１３１】関連非テキストコンテンツ判定部１０４は、判定対象文書Ｃが、辞書順で最も若いファイル名を持つと判定した場合（ステップＳ３９：Ｙｅｓ）、ステップＳ４０に進み、判定対象文書Ｃを非テキストコンテンツテーブル１１５に登録し、その文書についての処理を終了する。そうでない場合（ステップＳ３９：Ｎｏ）、ステップＳ４０を行わないで、その文書について

の処理を終了する。

【0132】例えば、リンク元文書Sがアルバムのように画像を一覧表示する内容の文書である場合、これらの全てをリンク元文書Sの内容に関連する文書として扱うと、関連する文書が多くなり、かえって利用者に検索結果を提供する際に煩雑となってしまうことが考えられる。しかし、このような場合、例えば、pict01.jpg、pict02.jpg、pict03.jpg、・・・のように、数値部分を除いた残りの部分は互いに同一である事が多い。従って、互いに類似したファイル名を持つリンク先文書がある場合、辞書順に最も若いファイル名を持つ文書のみを関連非テキストコンテンツとして登録することにより、このような煩雑さを避けることが可能となる。

【0133】上述のようにして、ある判定対象文書Cについての処理を終了した後、関連非テキストコンテンツ判定部104は、リンク元文書Sのリンク関係情報を参照し、先に取り出したリンク元文書Sに他の未判定のリンク先文書があるか否か判定する。未判定のリンク先文書がある場合、関連非テキストコンテンツ判定部104は、その未判定のリンク先文書を新たな判定対象文書Cとし、その文書についてステップS31以降の処理を行う。

【0134】また、そのリンク元文書Sに他の未判定のリンク先文書が含まれていない場合は、関連非テキストコンテンツ判定部104は、他の未処理のリンク元文書Sをリンク元文書集合から取り出して、そのリンク元文書Sのリンク先文書Cについて同様の処理を行う（不図示）。また、全てのリンク元文書Sについて処理を行った場合、関連非テキストコンテンツ判定処理を終了する。

【0135】各文書についての情報をユーザに提供する際に、その文書の文書位置情報、タイトル及び内容を示す情報とともに、上記判定結果に基づいてその文書からリンクされている関連非テキストコンテンツの種別を示す情報、例えばアイコンをユーザに提供することとしても良い。これにより、ユーザは、その文書のリンク先にどのような関連非テキストコンテンツがあるのか、その文書を実際に閲覧（ブラウズ）することなく知ることができる。また、さらに、上述の関連非テキストコンテンツの種別を示すアイコンに、その関連非コンテンツへのリンクを埋め込む事により、ユーザがアイコンを選択（クリック、或いはタッチ等）した場合に、その関連非テキストコンテンツをユーザの画面に表示又は再生等させることとしても良い（後述）。

【0136】次に、図13を用いて文書のサービス種別を判定する処理の手順について説明する。文書において、様々なサービスがその文書の閲覧者に提供されていることが多い。サービス種別判定部105は、文書中で用いられているフォームタグに基づいて、その文書で提供されているサービスの種別を判定する。以下の説明において、検索、ショップ及び申込（登録）の3つのサービス種別を判定している。

【0137】ここで、検索サービスとは、ユーザ（又は閲覧者等）が入力されたキーワードに基づいて何かを探すサービスをいう。ショップサービスとは、ユーザに商品を販売するサービスをいう。申込（登録）サービスとは、ユーザから氏名や住所等を受け付け、ユーザから会員や懸賞の申込又は登録を受け付けるサービスをいう。なお、これらの3つのサービスは、例示であり、本発明を限定する趣旨ではない。サービス種別を判定する処理に、さらに多くの手順を追加することによって、更に詳しくサービス種別を判定することが可能となる。

【0138】まず、サービス種別判定部105は、収集済みの文書のうちテキストが含まれる文書を抽出する（不図示）。テキストが含まれているか否かは、例えば各文書のファイル名の拡張子に基づいて判定する事にしても良い。以下の処理は、抽出された各文書について行われる。

【0139】続いて、サービス種別判定部105は、文書にフォームタグが含まれるか否か判定する（ステップS41）。文書にフォームタグが含まれない場合（ステップS41：No）、その文書はサービスを提供していないと推測されるため、その文書についての処理を終了する。

【0140】文書にフォームタグが含まれる場合（ステップS41：Yes）、サービス種別判定部105は、更に、その文書に含まれるボタンに「購入」又は「買う」等の文字があるか否か判定する（ステップS42）。

【0141】例えば、HTMLで記述された文書の場合、ボタンは以下のように表記されることが多い。

<;INPUT TYPE="submit" VALUE="ボタンに表示する文字">;

ボタンに「購入」、「purchase」又は「買う」等の文字がある場合（ステップS42：Yes）、サービス種別判定部105は、その文書で提供されるサービスの種別を「ショップ（販売店）」であると判定し（ステップS43）、ステップS48に進む。サービス種別判定部105は、その文書の文書IDとともに判定したサービス種別「ショップ」をサービス種別テーブル116に格納する事により、その文書のサービス種別を「ショップ」として登録する（ステップS48）。

【0142】ボタンに「購入」又は「買う」等の文字がない場合（ステップS42：No）、サービス種別判定部105は、更に、その文書にユーザの入力エリアが含まれるか否か判定する（ステップS44）。ユーザの入力エリアが含まれない場合（ステップS44：No）、その文書でサービスは提供されていないと推測し、その文書についての処理を終了する。その文書にユーザの入力エリアが含まれる場合（ステップS44：Yes）、サービス種別判定部105は、更に、その文書に含まれ

るボタンに「検索」又は「search」等の文字があるか否か判定する（ステップＳ４５）。

【０１４３】ボタンに「検索」又は「search」等の文字がある場合（ステップＳ４５：Ｙｅｓ）、サービス種別判定部１０５は、その文書が提供するサービスの種別を「検索」であると判定し（ステップＳ４６）、ステップＳ４８に進む。ステップＳ４８において、サービス種別判定部１０５は、上述のようにしてその文書が提供するサービスを登録する。

【０１４４】ボタンに「検索」又は「search」等の文字がない場合（ステップＳ４５：Ｎｏ）、サービス種別判定部１０５は、その文書が提供するサービスの種別を「申込」であると判定し（ステップＳ４７）、ステップＳ４８に進む。

【０１４５】このように、サービス種別判定部１０５は、文書の内容を見ることなく、フォームタグに基づいて、その文書で提供されているサービスの種別を判定することができる。

【０１４６】なお、サービス種別を判定する処理には、様々な変形が考えられる。例えば、ステップＳ４５とステップＳ４６の間で以下の処理を行う事としてもよい。まず、ステップＳ４５の後、サービス種別判定部１０５は、更に、ＩＳＢＮ（International Standard Book Number:国際標準図書番号）の入力欄があるか否か判定し、ＩＳＢＮの入力欄が含まれる場合、その文書が提供するサービスの種別を「書店」として判定してステップＳ４８に進む。ＩＳＢＮの入力欄が含まれない場合、ステップＳ４６に進む。これにより、文書が提供しているサービスを更に詳しく判定する事が可能となる。

【０１４７】各文書についての情報をユーザに提供する際に、その文書の文書位置情報、タイトル及び内容を示す情報とともに、上記判定結果に基づいて、その文書が提供するサービスの種別を示す情報、例えばアイコンをユーザに提供することとしても良い。これにより、ユーザは、その文書が提供しているサービスの種別をその文書を実際に閲覧（ブラウズ）することなく知ることができる。また、上記判定において判定されたサービス種別が、各ページを分類する際に使用する事ができる。

【０１４８】ページ分類部１０６は、各文書中の語句に基づいて、その文書の内容を判定し、判定結果に基づいて各文書を分類する。文書の内容を示す語句として、例えば、「Ｊａｖａ（登録商標）」、「テーマパーク」等が考えられる。なお、この例示は本発明を限定する趣旨ではない。このページ分類部による各文書の分類方法は従来技術と同じであるため、詳しい説明は省略する。なお、ページ分類部１０６は、各文書を分類する際に、例えば、サービス種別判定部１０５によって判定された各文書で提供されるサービス種別を利用することとしても良い。

【０１４９】検索サービス部１０７は、文書検索装置１００のユーザからの指示に基づいて文書を検索し、適宜上述の人気度算出部１０２及び人気度遷移算出部１０３等の処理結果とともに検索結果をそのユーザに対して提供する。より具体的は、検索サービス部１０７は、ユーザの端末に処理結果とともに検索結果を表示させる。以下、検索サービス部１０７が行う処理について、ユーザの端末に表示される画面を適宜参照しながら説明する。

【０１５０】検索サービス部１０７は、検索の結果得られた文書に関する情報を、さまざまな形式でユーザに提供する。まず、ユーザがキーワード等を入力し、そのキーワード等に基づいて検索した結果をユーザに提供する場合について説明する。

【０１５１】まず、検索サービス部１０７は、ユーザが入力したキーワード等に基づいて、文書を検索し、検索された文書について、以下の情報を各テーブルから取得する。

【０１５２】・最新の人気度及び人気度順位を人気度テーブル１１３から取得する。

・最新の人気度及び人気度順位のそれぞれに基づく回帰係数ａ（傾き）及び切片ｂを人気度変化テーブル１１４から取得する。

【０１５３】・関連非テキストコンテンツの文書ＩＤを非テキストコンテンツテーブル１１５から取得する。

・サービス種別をサービス種別テーブル１１６から取得する。

【０１５４】続いて、検索サービス部１０７は、取得した回帰係数ａ及び切片ｂに基づいて、人気度の変化の方向と速度を図示する人気度推移アイコンを作成する。人気度推移アイコンは、具体的には、矢印を図示するアイコンであり、人気度の変化の方向と速度を矢印の向きと傾きで示す。検索サービス部１０７は、人気度推移アイコンとして、例えば、以下の６種を作成する。なお、この例示は、本発明を限定する趣旨ではない。

【０１５５】急上昇アイコン：人気度が急激に上昇している事を示す。急上昇アイコンは、角度が急な右肩上がりの矢印を図示する。

上昇アイコン：人気度が上昇している事を示す。上昇アイコンは、右肩上がりの矢印を図示し、その角度は、急上昇アイコンよりも水平に近い。

【０１５６】下降アイコン：人気度が下降している事を示す。下降アイコンは、右肩下がりの矢印を図示し、その角度は、急下降アイコンよりも水平に近い。

急下降アイコン：人気度が急激に下降している事を示す。急上昇アイコンは、角度が急な右肩下がりの矢印を図示する。

【０１５７】安定アイコン：右向きの水平の矢印を図示する。後述の高値安定と低値安定の場合とで色を変えることとしてもよい。

無印アイコン：矢印がないアイコンである。その他の状態を示す。

【０１５８】人気度推移アイコンの作成方法の例として、以下の２つを挙げる。
（例１） 人気度変化を人気度（10000までの自然数。大きいほど人気度が高い）を元に計算した場合
検索サービス部１０７は、以下のようにして回帰係数ａ及び切片ｂに基づいて各文書に付すべきアイコンを判定する。
【０１５９】急上昇アイコン：その文書のａが５０以上の場合
上昇アイコン：その文書のａが３０以上の場合
下降アイコン：その文書のａが－３０以下の場合
急下降アイコン：その文書のａが－５０以下の場合
高値安定アイコン：その文書のｂが８０００以上の場合
低値安定アイコン：その文書のｂが３０００以下の場合
無印アイコン：その他の場合
（例２）人気度変化を人気度順位（１から総文書数までの自然数。小さいほど人気度順位がよい）で計算した場合
検索サービス部１０７は、以下のようにして各文書に付すべきアイコンを判定する。
【０１６０】急上昇アイコン：その文書のａ／ｂが－０．１以下（１０％以上増加）の場合
上昇アイコン：その文書のａ／ｂが－０．０５以下（５％以上増加）の場合
下降アイコン：その文書のａ／ｂ が０．０５以上（５％以上減少）の場合
急下降アイコン：その文書のａ／ｂ が０．１以上（１０％以上減少）の場合
高値安定：その文書のｂが１０００以下の場合
低値安定：その文書のｂが１００００以上の場合
無印　　：その他の場合
続いて、検索サービス部１０７は、関連非テキストコンテンツが登録されていた文書について、関連非テキストコンテンツの種類を図示する関連メディアアイコンを作成し、その関連メディアアイコンに関連非テキストコンテンツへのリンクを埋め込む。これにより、関連メディアアイコンをユーザが選択すると、その関連非テキストコンテンツのリンク元文書を閲覧することなく関連非テキストコンテンツを閲覧、再生等させることが可能となる。
【０１６１】関連メディアアイコンは、例えば、関連非テキストコンテンツの種別を表示する。より具体的には、関連非テキストコンテンツがｊｐｇ形式である場合、関連メディアアイコンは、「ｊｐｇ」という文字列を表記する。或いは、関連メディアアイコンは、画像を示すように、カメラを図示する事としても良い。なお、文書に複数の関連非テキストコンテンツが登録されている場合、各関連非テキストコンテンツについてこの処理を行う。
【０１６２】さらに、検索サービス部１０７は、サービス種別が登録されていた文書について、サービス種別の種類を図示するサービス内容アイコンを作成する。サービス内容アイコンは、例えば、サービスの種別を表示するアイコンである。より具体的には、サービス種別がショップである場合、サービス内容アイコンは、「ショップ」という文字列を表記する。或いは、サービス内容アイコンは、ショップを図示する事としても良い。
【０１６３】最後に、検索サービス部は、検索の結果得られた各文書を人気度順位に基づいてソートし、ソートした順に、各文書のタイトル、文書の内容を示す情報、文書の文書位置情報、人気度推移アイコン、関連メディアアイコン及びサービス内容アイコンを画面に設定する。これにより、図１４に示すような、検索結果の表示画面が作成される。
【０１６４】図１４に示す検索結果の表示画面において、各文書は、最新の人気度の順、つまり静的な人気度の順に並べられる。ユーザは、各文書の人気度がどのように変化した結果、この順位になったのか、人気度推移アイコンによって知ることができる。さらに、ユーザは、関連メディアアイコンによって、各文書はどのような非テキスト文書にリンクしているのか知ることができ、さらに、関連メディアアイコンを選択（クリック、或いはタッチ等）することにより、関連非テキストコンテンツを再生、又は閲覧等する事が可能である。従って、ユーザは、その文書を閲覧することなく、その文書からどのような非テキストコンテンツにリンクしているのかを知ることが可能となる。
【０１６５】また、さらに、ユーザは、サービス内容アイコンによって、各文書はどのようなサービスを提供しているのか知ることができる。図１４において、ユーザが人気度推移アイコンを選択（クリック、或いはタッチ等）すると、検索サービス部１０７は、人気度推移アイコンが選択された文書について、過去一定期間内、例えば各数ヶ月内に算出された人気度又は人気度順位を人気度テーブル１１３から取得し、人気度が算出された日付に対する人気度又は人気度順位のグラフを作成し、画面に設定する。
【０１６６】図１５（ａ）に、人気度が算出された日付に対する人気度順位のグラフが設定された人気度推移画面の一例を示す。図１５（ａ）において、横軸が日付、縦軸が人気度順位を示す。また、グラフ中において数字は上下に記載されているが、上段の数字は人気度順位を示し、下段の数字は人気度が算出された日付を示す。このグラフは、当該文書の人気度が、この数ヶ月どのように推移したのかを示したものであり。人気度変化テーブルを視覚化したものに相当する。図１５（ａ）に示すように、ＵＲＬ：ｗｗｗ．ａａａによって特定される文書の人気度順位は、３月に急上昇した後、５月以降ほぼ安定して推移している事が分かる。
【０１６７】図１５（ａ）において、グラフ中の一部が

選択されると、検索サービス部１０７は、その選択された付近の適当な期間内の日付を収集日又は更新日とし、その文書の文書ＩＤをリンク先ＩＤとするリンク関係情報をリンク関係テーブル１１２から取得する。そして、検索サービス部１０７は、取得したリンク関係情報に基づいて、その一定期間内にその文書をリンク先としていた文書の一覧を作成し、画面に設定する。

【０１６８】図１５（ｂ）にある期間内において、ＵＲＬ：ｗｗｗ．ａａａで特定される文書をリンク先としていた文書、つまり、ＵＲＬ：ｗｗｗ．ａａａで特定される文書のリンク元文書の一覧を示す画面の一例を示す。図１５（ｂ）によって、ユーザは、その時期に、その文書がどのような文書からリンクされているのか知ることができる。例えば、ユーザが、ＵＲＬ：ｗｗｗ．ａａａで特定される文書のサイトマスターである場合、ユーザは、今後のサイトのメンテナンスにこの情報を応用する事が可能となる。

【０１６９】また、更に、ユーザは、予めある文書の文書位置情報及び人気度の閾値を検索サービス部１０７に登録しておき、検索サービス部１０７は、その文書の人気度が閾値以上又は閾値以下になった場合に、そのユーザに通知する事としてもよい。この場合も、ユーザは、その文書の人気度の変化を自動的に知ることができるため、ユーザは、今後のサイトのメンテナンス等にこの情報を応用する事が可能となる。

【０１７０】また、本発明の文書検索装置は、一般的な検索以外のその他、様々な用途に利用可能である。例えば、文書検索装置１００を、業界分析ツールとして利用することもできる。文書検索装置１００を利用して特定業界の人気度推移を表示し、ユーザはこの人気度推移をマーケティングの助けにすることができる。そのために、利用者は、まず、知りたい業界の企業トップページ（文書）の文書位置情報の一覧（例えばＵＲＬ集）を作成する。

【０１７１】続いて、文書検索装置１００は、文書位置情報の一覧に含まれる各文書の最新の人気度を人気度テーブル１１３から取得し、取得した人気度が高い順に各文書を一覧表示した人気度リストを設定する。この人気度リストは現在の業界ランキングを意味する。

【０１７２】図１６（ａ）に、人気度リストの一例を示す。図１６（ａ）の下端に「過去１ヶ月」及び「過去１年」と表示されたボタンが設定されている。このボタンが押下されると、文書検索装置は、さらに、過去１ヶ月間又は過去１年間に算出された文書位置情報一覧に含まれる各文書の人気度を人気度テーブル１１１から取得し、人気度を算出した日付に対する人気度の推移を示すグラフを作成し、画面に設定する。なお、人気度の代わりに人気度順位を用いても良いことはいうまでもない。

【０１７３】図１６（ｂ）に、過去１年の各文書の人気度の推移を示すグラフの一例を示す。図１６（ｂ）は、図１６（ａ）に示すリスト内の各文書の過去１年の人気度の推移を示し、図１６（ａ）において「過去１年」と表記されたボタンが押下された場合に、ユーザの端末に表示される。図１６（ｂ）において、横軸は、人気度が算出された日付を、縦軸は人気度を示す。図１６（ｂ）に示すように、ＵＲＬ：ｂｂｂ．ｃｏ．ｊｐを持つ文書の人気度が過去１年で急上昇している事が分かる。

【０１７４】また、例えば、文書検索装置１００を、地域情報検索システムとして利用する事も可能である。そのために、まず、ページ分類部１０６は、例えば、都道府県、市町村等のような地域を示す階層的なカテゴリを作成し、そのカテゴリに従って各文書を分類する。ユーザは、階層的なカテゴリを辿って、求める文書とその人気度、人気度推移、参照メディア、ページで提供するサービスにアクセスさせることができる。

【０１７５】図１７に、地域情報検索システムの画面の一例を示す。図１７（ａ）に、カテゴリ「東京都」に関する文書を一覧表示する画面の一例を示す。図１７（ａ）において、画面の上段に選択された地域「東京都」が表示され、中段に東京都内の各区が表示され、下段に「東京都」に分類された各文書に関する情報が表示されている。画面の下段は、図１４に示す検索結果の表示画面と同様であるため、図１７において省略している。図１７（ａ）の画面の上段においてユーザが「港区」を選択すると、カテゴリ「港区」に関する文書を一覧表示する画面に遷移する。

【０１７６】図１７（ｂ）に、カテゴリ「東京都－港区」に関する文書を一覧表示する画面の一例を示す。図１７（ｂ）において、画面の上段に選択された地域「港区」が表示され、画面の中段に港区内の町名が表示され、画面の下段に「東京都－港区」に分類された各文書に関する情報が表示されている。画面の下段は、図１４に示す検索結果の表示画面と同様である。図１７（ｂ）の画面の上段においてユーザが更に「六本木」を選択すると、カテゴリ「東京都－港区－六本木」に関する文書を一覧表示する画面に遷移する。

【０１７７】図１７（ｃ）に、カテゴリ「東京－港区－六本木」に関する文書を一覧表示する画面の一例を示す。図１７（ｃ）において、画面の上段に選択された地域「六本木」が表示され、画面の中段にその他のカテゴリが表示され、画面の下段に「東京都－港区－六本木」に分類された文書に関する情報が表示されている。

【０１７８】本実施形態において説明した文書検索装置１００及びユーザの端末等は、図１８に示すようなコンピュータ（情報処理装置）を用いて構成することもできる。図１８のコンピュータ２００は、ＣＰＵ２０１、メモリ２０２、入力装置２０３、出力装置２０４、外部記憶装置２０５、媒体駆動装置２０６、及びネットワーク接続装置２０７を備え、それらはバス２０８により互いに接続されている。

【０１７９】メモリ２０２は、例えば、ＲＯＭ（ReadOnly Memory）、ＲＡＭ（Random Access Memory）等を含み、処理に用いられるプログラムとデータを格納する。ＣＰＵ２０１は、メモリ２０２を利用してプログラムを実行することにより、必要な処理を行う。

【０１８０】コンピュータ２００に文書検索装置１００に相当する機能を実現させる場合、図１に示す文書検索装置１００を構成する収集部１０１、人気度算出部１０２、人気度遷移算出部１０３、関連非テキストコンテンツ判定部１０４、サービス種別判定部１０５、ページ分類部１０６及び検索サービス部１０７は、各部によって行われる処理を示すプログラムとして実現され、それぞれメモリ２０２の特定のプログラムコードセグメントに格納される。なお、上述の各部によって行われる処理は、各フローチャートにおいて説明されている。

【０１８１】入力装置２０３は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、ユーザからの指示や情報の入力に用いられる。出力装置２０４は、例えば、ディスプレイやプリンタ等であり、コンピュータ２００の利用者への問い合わせ、処理結果等の出力に用いられる。

【０１８２】外部記憶装置２０５は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク装置等である。この外部記憶装置２０５に上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ２０２にロードして使用することもできる。

【０１８３】媒体駆動装置２０６は、可搬出記録媒体２０９を駆動し、その記録内容にアクセスする。可搬出記録媒体２０９としては、メモリカード、メモリスティック、フレキシブルディスク、ＣＤ－ＲＯＭ（Compact Disc Read Only Memory）、光ディスク、光磁気ディスク、ＤＶＤ（Digital VersatileDisk）等、任意のコンピュータで読み取り可能な記録媒体が用いられる。この可搬出記録媒体２０９に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ２０２にロードして使用することもできる。

【０１８４】ネットワーク接続装置２０７は、ＬＡＮ、ＷＡＮ等の任意のネットワーク（回線）を介して外部の装置を通信し、通信に伴なうデータ変換を行う。また、必要に応じて、上述のプログラムとデータを外部の装置から受け取り、それらをメモリ２０２にロードして使用することもできる。

【０１８５】図１９は、図１８のコンピュータにプログラムとデータを供給することができる、コンピュータで読み取り可能な記録媒体及び伝送信号を説明する図である。上述のプログラムや各テーブルに格納されるデータを、以下のようにしてコンピュータ２００に供給することにより、コンピュータ２００に文書検索装置１００に相当する機能を行なわせることも可能である。そのためには、上述のプログラムやデータを、コンピュータで読み取り可能な記録媒体２９に予め記憶させておく。そして、図１９に示すように、媒体駆動装置２０６を用いて、記録媒体２９からプログラム等をコンピュータ２００に読み出させて該コンピュータ２００のメモリ２０２や外部記憶装置２０５に一旦格納させ、そのコンピュータ２００の有するＣＰＵ２０１にこの格納されたプログラムを読み出させて実行させるように構成すればよい。

【０１８６】また、記録媒体２０９からプログラムをコンピュータに読み出させる代わりに、プログラム（データ）提供者が有するＤＢ２１０から、通信回線（ネットワーク）２１１を介して、プログラムをダウンロードすることとしてもよい。この場合、例えば、ＤＢ２１０を有しプログラムを送信するコンピュータでは、上記プログラムを表現するプログラム・データをプログラム・データ・シグナルに変換し、変換されたプログラム・データ・シグナルをモデムを用いて変調することにより伝送信号を得て、得られた伝送信号を通信回線２１１（伝送媒体）に出力する。プログラムを受信するコンピュータでは、受信した伝送信号をモデムを用いて復調することにより、プログラム・データ・シグナルを得て、得られたプログラム・データ・シグナルを変換することにより、プログラム・データを得る。

【０１８７】なお、送信側のコンピュータと受信側のコンピュータの間を接続する通信回線２１１（伝送媒体）がデジタル回線の場合、プログラム・データ・シグナルを通信することも可能である。また、データベース（ＤＢ）２１０を有し、プログラムを送信するコンピュータと、プログラムをダウンロードするコンピュータとの間に、電話局等のコンピュータが介在しても良い。

【０１８８】以上、本発明の実施形態について説明したが、本発明は上述した実施形態に限定されるものではなく、他の様々な変更が可能である。

（付記１）　ネットワーク上の文書の人気の高さの度合いである人気度を算出する人気度算出方法であって、文書からリンク関係を抽出し、第１の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出する、ことを含むことを特徴とする人気度算出方法。

【０１８９】（付記２）　前記リンク関係及び前記文書の前記ネットワーク上の位置を示す文書位置情報に基づいて前記人気度を算出する、ことを更に含むことを特徴とする付記１に記載の人気度算出方法。

【０１９０】（付記３）　前記文書位置情報を示す文字列の特徴に基づいて、前記人気度を算出する、ことを更に含むことを特徴とする付記２に記載の人気度算出方法。

【０１９１】（付記４）　前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出する、ことを更に含むことを特徴とする付記１に記載の人気度算出方法。

【0192】（付記5） 第2の期間内に算出された前記人気度に基づいて、前記人気変化度を算出する、ことを更に含むことを特徴とする付記4に記載の人気度算出方法。

【0193】（付記6） 前記第2の期間内に算出された前記人気度の時間に対する回帰式を算出し、前記人気変化度を前記回帰式に基づいて算出する、ことを更に含むことを特徴とする付記5に記載の人気度算出方法。

【0194】（付記7） 前記回帰式の回帰係数に基づいて前記人気変化度を決定する、ことを更に含むことを特徴とする付記6に記載の人気度算出方法。

（付記8） 前記回帰式の切片に基づいて、前記人気度の時間に対する推移の傾向を決定する、ことを更に含むことを特徴とする付記7に記載の人気度算出方法。

【0195】（付記9） 前記第2の期間内に算出された前記人気度に基づいて、前記抽出された文書中の各文書の順位を決定し、前記第2の期間内の前記順位の時間に対する回帰式を算出し、前記人気変化度を前記回帰式に基づいて算出する、ことを更に含むことを特徴とする付記5に記載の人気度算出方法。

【0196】（付記10） ネットワーク上の文書間の関係を判定する文書関係判定方法であって、第1の文書からリンク関係を抽出し、前記リンク関係に基づいて、前記第1の文書からリンクされる第2の文書が、前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを含むことを特徴とする文書関係判定方法。

【0197】（付記11） 前記第1の文書から前記第2の文書にリンクする部分の近辺にある文字列を前記第1の文書から抽出し、前記文字列に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記10記載の文書関係判定方法。

【0198】（付記12） 前記文字列が特定の文字列である場合、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であると決定する、ことを更に含むことを特徴とする付記11に記載の文書関係判定方法。

【0199】（付記13） 前記第2の文書のファイル名の拡張子に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0200】（付記14） 前記拡張子が特定の拡張子でない場合、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書でないと決定する、ことを更に含むことを特徴とする付記13に記載の文書関係判定方法。

【0201】（付記15） 前記第2の文書が前記第1の文書内で所定回数以上使用されているか否かに基づいて、前記第2の文書は前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0202】（付記16） 前記第2の文書が前記第1の文書内で所定回数以上使用されている場合、前記第2の文書は前記第1の文書の内容に関連する非テキスト文書でないと決定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0203】（付記17） 前記第2の文書が前記第1の文書内で所定回数以上使用されていない場合、前記第2の文書は前記第1の文書の内容に関連する非テキスト文書であると決定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0204】（付記18） 前記第1の文書内に前記第2の文書のファイル名と類似したファイル名を持つ第3の文書がある場合、前記第2の文書の前記ファイル名が前記第3の文書の前記ファイル名よりも辞書順に若くない場合、前記第2の文書を第1の文書の内容に関連する非テキスト文書としてデータベースに登録しない、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0205】（付記19） 前記第2の文書からリンクされる第3の文書があるか否か判定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0206】（付記20） 前記第2の文書からリンクされる第3の文書がある場合、前記第1の文書の前記ネットワーク上の位置を示す文書位置情報と前記第2の文書の文書位置情報に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記19に記載の文書関係判定方法。

【0207】（付記21） 前記第1の文書の前記文書位置情報と前記第3の文書の文書位置情報に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記20に記載の文書関係判定方法。

【0208】（付記22） 前記第2の文書の前記文書位置情報と第3の文書の前記文書位置情報が、前記第1の文書の前記文書位置情報と同じサーバアドレス又はドメインを持たない場合、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書でないと決定する、ことを更に含むことを特徴とする付記21に記載の文書関係判定方法。

【0209】（付記23） ネットワーク上の文書が提供するサービスの種別を判定するサービス種別判定方法であって、前記文書からユーザ入力を指定するタグを抽出し、前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定する、ことを含むことを特徴とするサービス種別判定方法。

【0210】（付記24） 前記文書に前記ユーザ入力を指定するタグが含まれていない場合、前記文書はサービスを提供しないと決定する、ことを更に含むことを特徴とする付記23に記載のサービス種別判定方法。

【0211】（付記25） 前記文書に含まれるボタンの表示に基づいて、前記文書が提供するサービスの種別を判定する、ことを更に含むことを特徴とする付記23に記載のサービス種別判定方法。

【0212】（付記26） 前記文書に含まれるユーザ入力エリアに基づいて、前記文書が提供するサービスの種別を判定する、ことを更に含むことを特徴とする付記25に記載のサービス種別判定方法。

【0213】（付記27） ネットワーク上の文書の人気の高さの度合いである人気度を算出する制御をコンピュータに実行させるプログラムであって、文書からリンク関係を抽出し、第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出する、ことを含む処理を前記コンピュータに実行させることを特徴とするプログラム。

【0214】（付記28） 前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出する、ことを更に含む処理を更にコンピュータに実行させることを特徴とする付記27に記載のプログラム。

【0215】（付記29） 第2の期間内に算出された前記人気度に基づいて、前記人気変化度を算出する、ことを更に含む処理を前記コンピュータに実行させる事を特徴とする付記28に記載のプログラム。

【0216】（付記30） 前記第2の期間内に算出された前記人気度の時間に対する回帰式を算出し、前記人気変化度を前記回帰式に基づいて算出する、ことを更に含む処理を前記コンピュータに実行させる事を特徴とする付記29に記載のプログラム。

【0217】（付記31） 前記回帰式の回帰係数に基づいて前記人気変化度を決定する、ことを更に含む処理を前記コンピュータに実行させる事を特徴とする付記30に記載のプログラム。

【0218】（付記32） 前記回帰式の切片に基づいて、前記人気度の時間に対する推移の傾向を決定する、ことを更に含む処理を前記コンピュータに実行させる事を特徴とする付記31に記載のプログラム。

【0219】（付記33） ネットワーク上の文書間の関係を判定する制御をコンピュータに実行させるプログラムであって、第1の文書からリンク関係を抽出し、前記リンク関係に基づいて、前記第1の文書からリンクされる第2の文書が、前記第1の文書の内容に関連する非テキストコンテンツであるか否か判定する、ことを含む処理を前記コンピュータに実行させる事を特徴とするプログラム。

【0220】（付記34） ネットワーク上の文書が提供するサービスの種別を判定する制御をコンピュータに実行させるプログラムであって、前記文書からユーザ入力を指定するタグを抽出し、前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定する、ことを含む処理を前記コンピュータに実行させる事を特徴とするプログラム。

【0221】（付記35） ネットワーク上から文書を検索する文書検索方法であって、前記ネットワークから文書を収集し、前記文書からリンク関係を抽出し、第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出し、検索条件に基づいて文書を検索し、前記検索された文書を前記人気度に基づいてランキングし、前記ランキング結果に基づいて、前記検索された文書に関する情報を出力する、ことを含むことを特徴とする文書検索方法。

【0222】（付記36） 第2の期間内に算出された前記人気度に基づいて、前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出し、前記人気変化度に関する情報を前記検索された文書に関連する情報に加える、ことを更に含むことを特徴とする付記35に記載の文書検索方法。

【0223】（付記37） 前記リンク関係に基づいて、前記文書からリンクされる他の文書が、前記文書の内容に関連する関連非テキスト文書であるか否か判定し、前記判定の結果に基づいて、前記関連非テキスト文書に関する情報を前記検索された文書に関連する情報に加える、ことを更に含むことを特徴とする付記35に記載の文書検索方法。

【0224】（付記38） 前記関連非テキスト文書に関する情報に、前記関連非テキスト文書へのリンクを埋め込む、ことを更に含むことを特徴とする付記37に記載の文書検索方法。

【0225】（付記39） 前記文書からユーザ入力を指定するタグを抽出し、前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定し、前記サービスの種別に関する情報を前記検索された文書に関連する情報に加える、ことを更に含むことを特徴とする付記35に記載の文書検索方法。

【0226】（付記40） ユーザからある文書の前記ネットワーク上の位置を示す文書位置情報及び所定値の登録を受け付け、前記文書位置情報によって特定される前記文書の前記人気度が、前記所定値になった場合、前記人気度が前記所定値になった旨を前記ユーザに通知する、ことを更に含むことを特徴とする付記35に記載の文書検索方法。

【0227】（付記41） ネットワーク上から文書を検索する文書検索装置であって、前記ネットワークから文書を収集し、前記収集された文書からリンク関係を抽出する収集手段と、第1の期間内に更新又は収集された

文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出する人気度算出手段と、検索条件に基づいて文書を検索し、前記検索された文書を前記人気度に基づいてランキングし、前記ランキング結果に基づいて、前記検索された文書に関する情報を出力する検索サービス手段と、を備えることを特徴とする文書検索装置。

【0228】（付記42） 地域に関する文書をネットワーク上から検索する地域情報文書検索装置であって、前記ネットワークから文書を収集し、前記収集された文書からリンク関係を抽出する収集手段と、第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出する人気度算出手段と、第2の期間内に算出された前記人気度に基づいて、前記人気度の変化の方向と度合いを示す人気変化度を算出する人気度遷移算出手段と、前記収集された文書間のリンク関係に基づいて、各文書からリンクされる文書が、各文書の内容に関連する関連非テキスト文書であるか否か判定する関連非テキストコンテンツ判定手段と、前記収集された文書からユーザ入力を指定するタグを抽出し、前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定するサービス種別判定手段と、前記収集された文書を地域名毎に階層的に分類する分類手段と、ユーザから指定された地域名に基づいて文書を検索し、前記検索された文書を前記人気度に基づいてランキングし、前記ランキング結果に基づいて、前記検索された文書に関する情報とともに、前記検索された文書の前記人気変化度に関する情報、前記関連非テキスト文書に関する情報及び前記検索された文書が提供するサービス種別に関する情報を出力する検索サービス手段と、を備えることを特徴とする文書検索装置。

【0229】

【発明の効果】以上詳細に説明したように、本発明は、第1の期間内に収集又は更新された文書を対象として人気の高さの度合いを示す人気度を算出し、さらに、第2の期間内に算出された人気度に基づいて人気度の変化の度合いを示す人気変化度を算出する。これにより、文書の人気度が増加する一方で減少することがないという問題を解決しつつ、文書が時系列的にどのような状況にあるのかを示す情報を得る事を可能とする。

【0230】また、本発明によれば、文書間のリンク関係及びタグに基づいて、非テキストコンテンツ及びサービスを提供する文書等、多様な文書を整理する事が可能となる。

【図面の簡単な説明】

【図1】本発明の原理図である。

【図2】本発明に係わる文書検索装置の構成図である。

【図3】文書テーブルのデータ構造の一例を示す図である。

【図4】リンク関係テーブルのデータ構造の一例を示す図である。

【図5】人気度テーブルのデータ構造の一例を示す図である。

【図6】人気度変化テーブルのデータ構造の一例を示す図である。

【図7】非テキストコンテンツテーブルのデータ構造の一例を示す図である。

【図8】サービス種別テーブルのデータ構造の一例を示す図である。

【図9】人気度を算出する処理の手順を示すフローチャートである。

【図10】人気度の算出における本発明の特徴を説明する図である。

【図11】人気変化度を算出する処理の手順を示すフローチャートである。

【図12】関連する非テキストコンテンツを判定する処理の手順を示すフローチャートである。

【図13】提供するサービスを判定する処理の手順を示すフローチャートである。

【図14】検索結果の表示画面の一例を示す図である。

【図15】人気度推移画面の一例を示す図である。

【図16】本発明を適用した業界分析ツールの画面の一例を示す図である。

【図17】本発明を適用した地域情報検索システムの画面の一例を示す図である。

【図18】コンピュータの構成図である。

【図19】コンピュータにプログラムやデータを提供することができる記録媒体及び伝送信号を説明する図である。

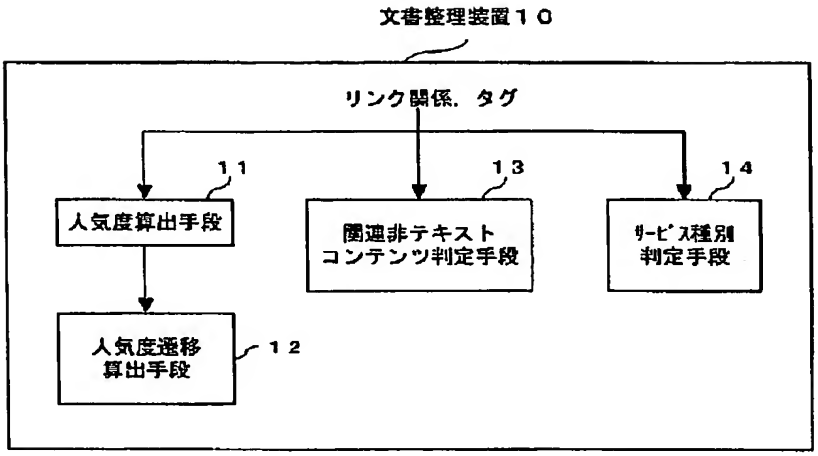
【符号の説明】

10 文書整理装置
11 人気度算出手段
12 人気度遷移算出手段
13 関連非テキストコンテンツ判定手段
14 サービス種別判定手段
100 文書検索装置
101 収集部
102 人気度算出部
103 人気度遷移算出部
104 関連非テキストコンテンツ判定部
105 サービス種別判定部
106 ページ分類部
107 検索サービス部
108 ブラウザ
111 文書テーブル
112 リンク関係テーブル
113 人気度テーブル
114 人気度変化テーブル
115 非テキストコンテンツテーブル

116　サービス種別テーブル
200　コンピュータ
201　CPU
202　メモリ
203　入力装置
204　出力装置
205　外部記憶装置
206　媒体駆動装置
207　ネットワーク接続装置
208　バス
209　可搬記録媒体
210　プログラム（データ）提供者
211　回線

【図1】

本　発　明　の　原　理　図



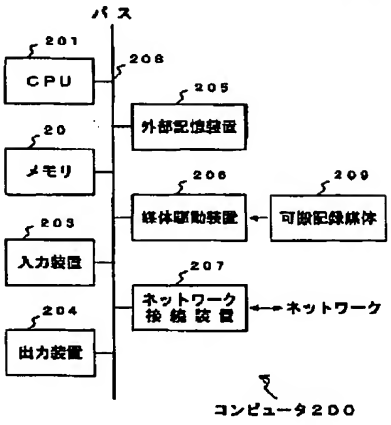
【図3】

文書テーブルのデータ構造の一例を示す図

文書テーブル111	
URL	ID
http://aaa.co.jp/	123
http://bbb.co.jp/dd/	124
.....	.....
.....	.....

【図18】

コンピュータの構成図



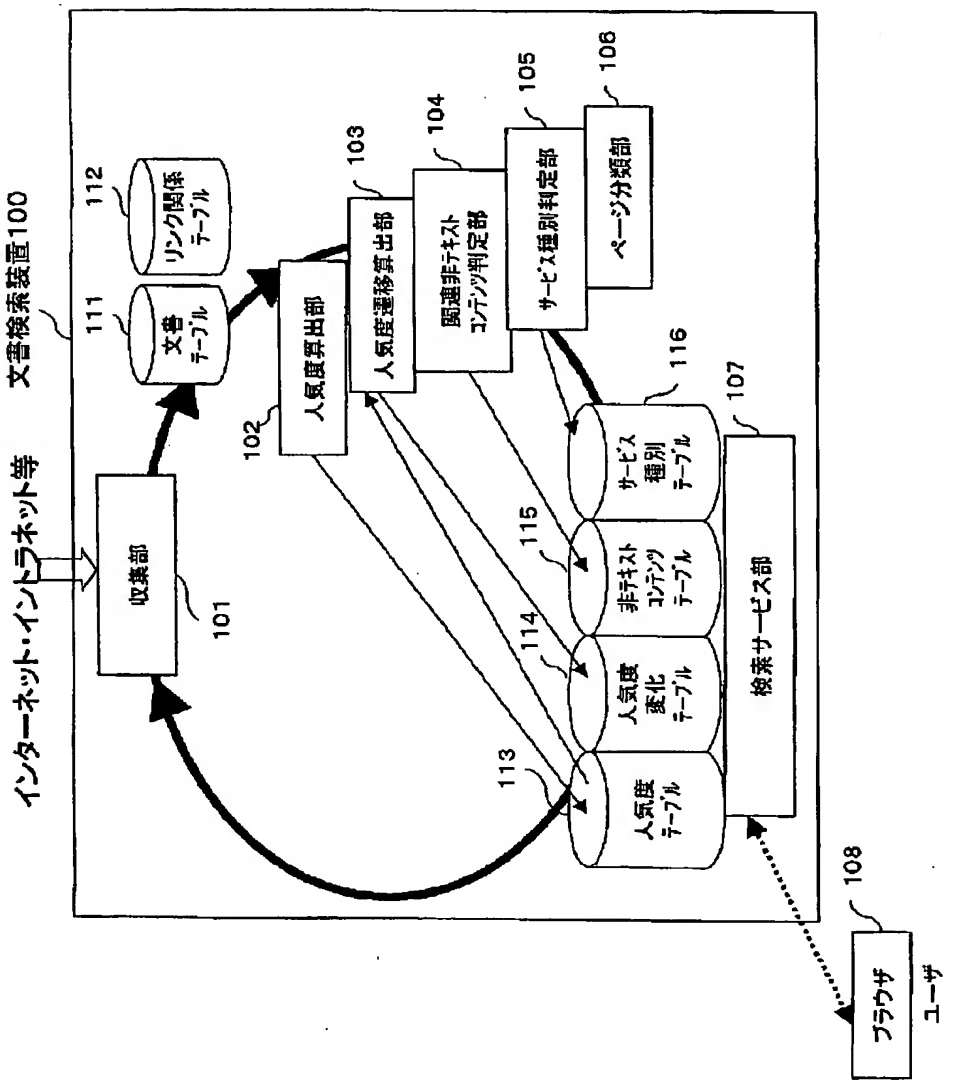
【図7】

非テキストコンテンツテーブルのデータ構造の一例を示す図

非テキストコンテンツテーブル115		
文書ID	関連非テキストコンテンツID	種別
123	3630	mv
123	3150	snd
.....	.....	.....

【図2】

本発明に係わる文書検索装置の構成図



【図4】

リンク関係テーブルのデータ構造の一例を示す図

リンク関係テーブル112

収集日	更新日	リンク元ID	リンク先ID列
010810	010725	123	124,128,3150,3630,…
010810	010620	124	256,975,1225,…..
….	…..	…..	…

【図5】

人気度テーブルのデータ構造の一例を示す図

人気度テーブル113

算出日	文書ID	人気度	人気度順位
010820	123	5036	346
010820	124	83645	5890
…..	…	…	…..

【図6】

人気度変化テーブルのデータ構造の一例を示す図

人気度変化テーブル114

文書ID	人気度		人気度順位	
	傾き	切片	傾き	切片
123	-12	346	-6	233
124	-562	5890	-152	851
…..	…	…		…..

【図8】

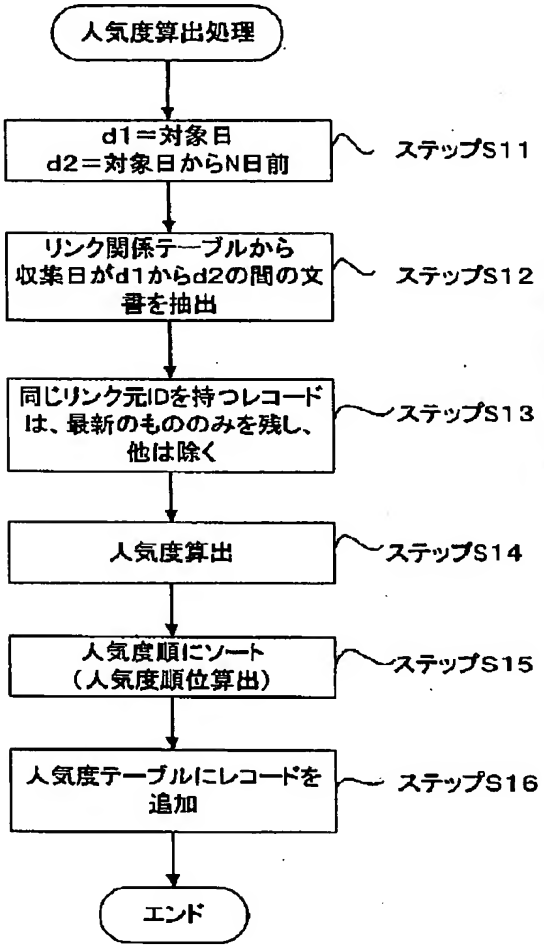
サービス種別テーブルのデータ構造の一例を示す図

サービス種別テーブル116

文書ID	サービス種別
124	検索
123	ショップ

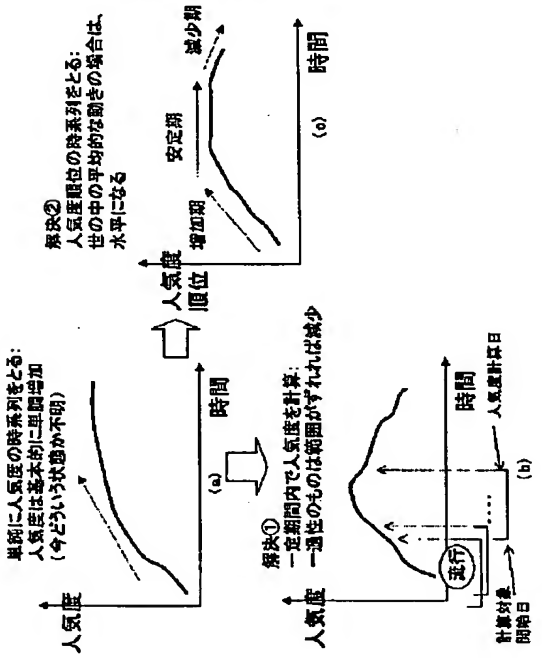
【図9】

人気度を算出する処理の手順を示すフローチャート



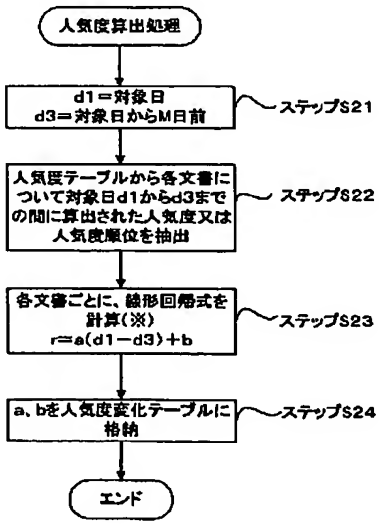
【図10】

人気度計算における本発明の特徴を説明する図



【図11】

人気変化度を算出する処理の手順を示すフローチャート

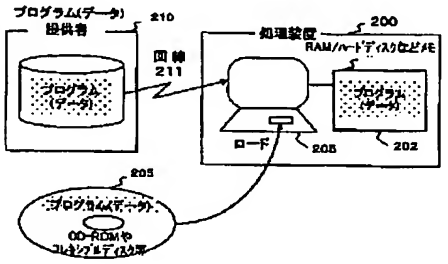


※d3、d3+1、・・・d1(M日)の人気度(順位)をw_0、w_1、・・・、w_{M-1}とすると、
$a=(M\times Iw-I\times W)/(M\times I2-I^2)$
$b=(I\times Iw-W\times I2)/(I^2-M\times I2)$である。
ただし、

$$Iw=\sum_{i=0}^{M-1} i*w_i \quad , \quad W=\sum_{i=0}^{M-1} w_i \quad , \quad I=\sum_{i=0}^{M-1} i \quad , \quad I2=\sum_{i=0}^{M-1} i^2$$

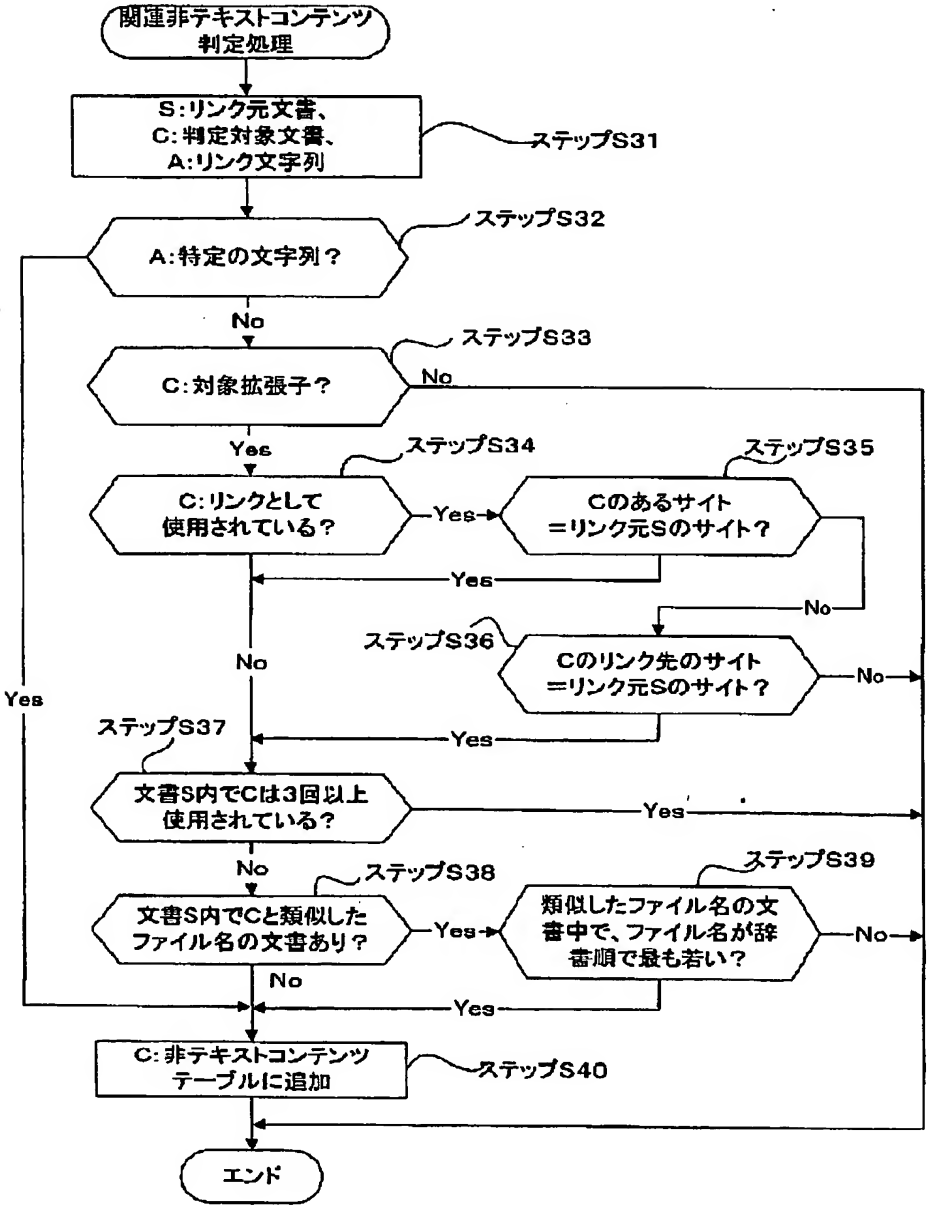
【図19】

コンピュータにプログラムやデータを提供することができる
記憶媒体及び伝送信号を説明する図



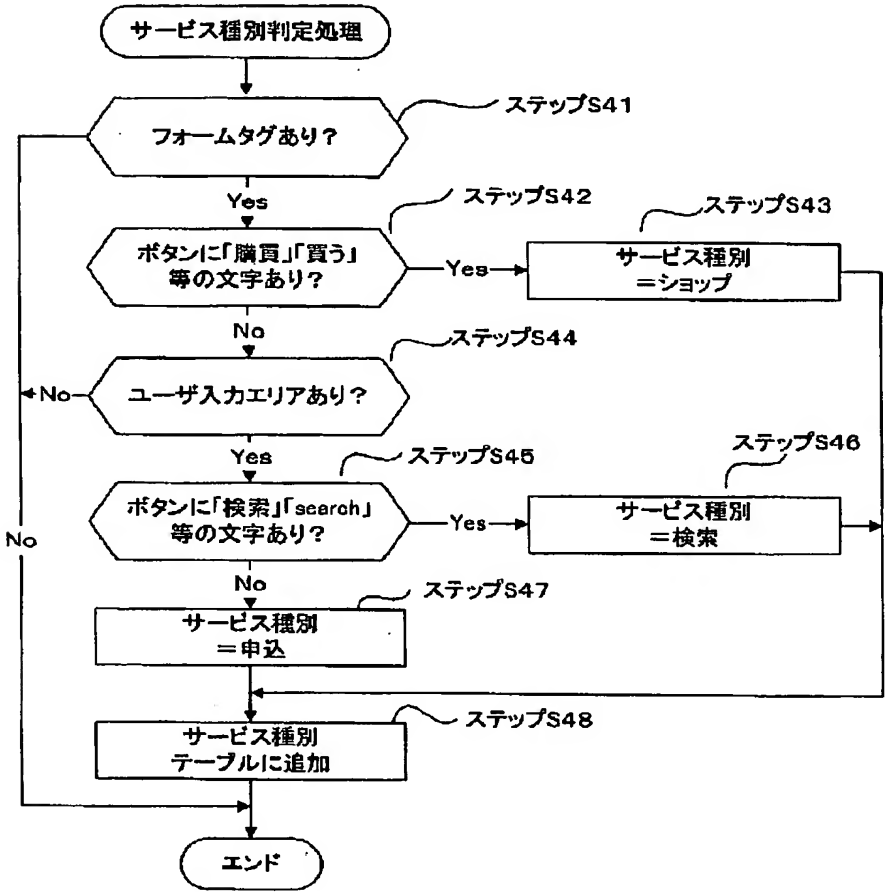
【図12】

関連する非テキストコンテンツを判定する処理の手順を示すフローチャート



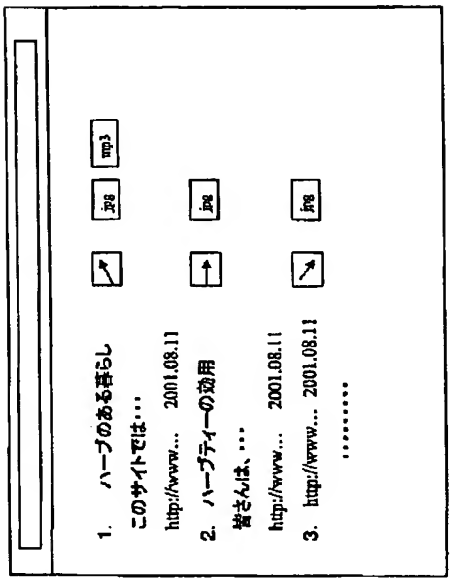
【図13】

提供するサービスを判定する処理の手順を示すフローチャート



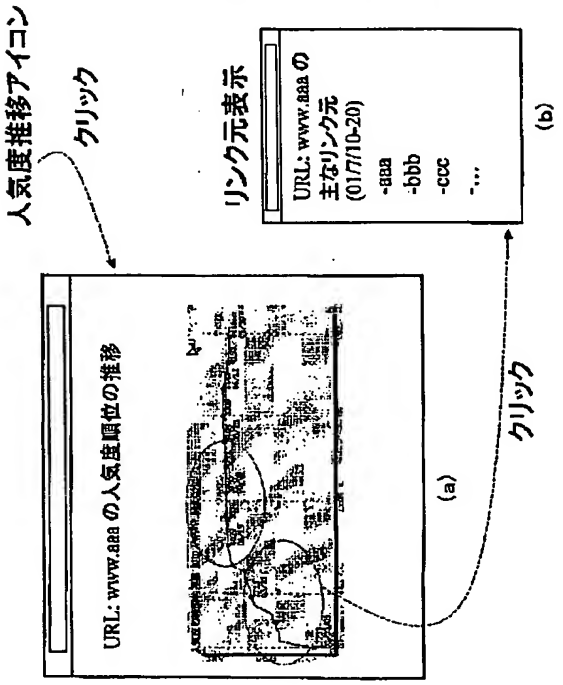
【図14】

検索結果の表示画面の一例を示す図



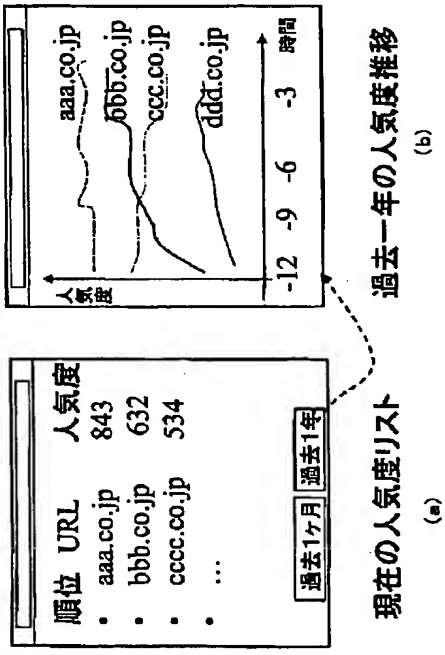
【図15】

人気度推移画面の一例を示す図



【図16】

本発明を適用した業界分析ツールの画面の一例を示す図



【図17】

本発明を適用した地域情報検索システムの画面の一例を示す図